Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Mixture-based clustering for the ordered stereotype model

D. Fernández[*], R. Arnold, S. Pledger

School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, Wellington, New Zealand

HIGHLIGHTS

- New methodology for clustering rows and columns from a matrix of ordinal data.
- Establishes likelihood-based methods via finite mixtures with the stereotype model.
- Tests the reliability of this methodology through a simulation study.
- Illustrates this new approach with two examples.
- Reviews and compares the performance several model choice measures.

ARTICLE INFO

Article history:
Received 20 March 2014
Received in revised form 6 November 2014
Accepted 8 November 2014
Available online xxxx

Keywords:
Biclustering
Cluster analysis
Dimension reduction
EM-algorithm
Finite mixture model
Fuzzy clustering
Likert scale
Ordinal data
Stereotype model

ABSTRACT

Many of the methods which deal with the reduction of dimensionality in matrices of data are based on mathematical techniques such as distance-based algorithms or matrix decomposition and eigenvalues. Recently a group of likelihood-based finite mixture models for a data matrix with binary or count data, using basic Bernoulli or Poisson building blocks has been developed. This is extended and establishes likelihood-based multivariate methods for a data matrix with ordinal data which applies fuzzy clustering via finite mixtures to the ordered stereotype model. Model-fitting is performed using the expectation–maximization (EM) algorithm, and a fuzzy allocation of rows, columns, and rows and columns simultaneously to corresponding clusters is obtained. A simulation study is presented which includes a variety of scenarios in order to test the reliability of the proposed model. Finally, the results of the application of the model in two real data sets are shown.

1. Introduction

An ordinal variable is one with a categorical data scale which describes order, and where the distinct levels of such a variable differ in degree of dissimilarity more than in quality (Agresti, 2010). This is different from nominal variables which vary in quality, not in quantity, and thus the order of listing the categories is irrelevant. For example, Likert scale responses in a questionnaire might be "disagree", "neither agree nor disagree" or "agree". In his seminal paper, Stevens (1946) called a scale ordinal if "any order-preserving transformation will leave the scale form invariant". Although the collection and use of ordinal variables is common, most of the current methods for analyzing them treat the data as if they were nominal (Hoffman and Franke, 1986) or continuous data (Agresti, 2010). On the one hand, treating an ordered categorical variable as ordinal rather than nominal provides advantages in the analysis such as simplifying the data description and allowing the use of more parsimonious models. The nominal approach ignores the intrinsic ordering of the data and thus the statistical

[*] Correspondence to: MSOR, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand. Tel.: +64 4 463 5341.
E-mail addresses: daniel.fernandez@msor.vuw.ac.nz (D. Fernández), richard.arnold@msor.vuw.ac.nz (R. Arnold), shirley.pledger@vuw.ac.nz (S. Pledger).

http://dx.doi.org/10.1016/j.csda.2014.11.004

results are less powerful than they could be. On the other hand, models for continuous variables have similarities to those for ordinal variables although the use of them with ordinal variables has disadvantages such as the treatment of the output categories as equally spaced, which they may not be (see Agresti, 2010, Sections 1.2–1.3 for a list of advantages from treating an ordinal variable as ordinal rather than nominal or continuous).

Categorical data analysis methods were first developed in the 1960s and 1970s (Bock and Jones, 1968; Snell, 1964), including loglinear models and logistic regression (see the review by Liu and Agresti, 2005). An increasing interest in ordinal data has since produced the articles by Goodman (1979) and McCullagh (1980) on loglinear modeling relating to ordinal odds ratios, and logit modeling of cumulative probabilities respectively. Recently, new ordinal data analysis methods have been introduced such as the proportional odds model version of the cumulative logit model, and the stereotype model with ordinal scores (Agresti, 2010, Chap. 3 and 4) from which new lines of research have developed. Two recent examples of these are the application of a stereotype model in a case-control study by Ahn et al. (2009), and a new methodology to fit a stratified proportional odds model by Mukherjee et al. (2008). In particular, the stereotype model is a paired-category logit model which is an alternative when the fit of cumulative logits and adjacent-categories logit models in their proportional odds version is poor. Anderson (1984) proposed this model as nested between the adjacent-categories logit model and the standard baseline-category logits model (see the review by Agresti, 2002, Chapter 6).

In the research literature, multiple algorithms and techniques have been developed which deal with the clustering of data such as hierarchical clustering (Johnson, 1967; Kaufman and Rousseeuw, 1990), association analysis (Manly, 2005) and partition optimization methods such as the k-means clustering algorithm (Jobson, 1992; Lewis et al., 2003; McCune and Grace, 2002). There has been research on cluster analysis for ordinal data based on latent class models (see Agresti and Lang, 1993; Moustaki, 2000; Vermunt, 2001; DeSantis et al., 2008; Breen and Luijkx, 2010; McPartland and Gormley, 2013 and the review by Agresti, 2010, Section 10.1). There are a number of clustering methods based on mathematical techniques such as distance metrics (Everitt et al., 2001), association indices (Wu et al., 2008; Chen et al., 2011), matrix decomposition and eigenvalues (Quinn and Keough, 2002; Manly, 2005; Wu et al., 2007). However, these do not have a likelihood based formulation, and do not provide a reliable method of model selection or assessment. A particularly powerful model-based approach to one-mode clustering based on finite mixtures, with the variables in the columns being utilized to cluster the subjects in the rows, is provided by McLachlan and Basford (1988), McLachlan and Peel (2000), Everitt et al. (2001), Böhning et al. (2007), Wu et al. (2008) and Melnykov and Maitra (2010).

The simultaneous clustering of rows and columns into row clusters and column clusters is called biclustering (or block clustering or two-mode clustering). Biclustering models based on double k-means have been developed in Vichi (2001) and Rocci and Vichi (2008). A hierarchical Bayesian procedure for biclustering is given in DeSarbo et al. (2004). Biclustering using mixtures has been proposed for binary data in Pledger (2000), Arnold et al. (2010) and Labiod and Nadif (2011), and for count data in Govaert and Nadif (2010). An approach via finite mixtures for binary and count data using basic Bernoulli or Poisson building blocks has been developed in Govaert and Nadif (2010) and Pledger and Arnold (2014). This work expanded previous research for one-mode fuzzy cluster analysis based on finite mixtures (McLachlan and Basford, 1988; McLachlan and Peel, 2000; Everitt et al., 2001) to a suite of models including biclustering. Finally, Matechou et al. (2011) have recently developed biclustering models for ordinal data using the assumption of proportional odds and having a likelihood-based foundation. The main difference with our work is that we use the assumption of ordinal stereotype model which has the advantage of allowing us to determine a new spacing of the ordinal categories, dictated by the data.

In this article, we present an extension of the likelihood-based models proposed in Pledger and Arnold (2014) by applying them to matrices with ordinal data by using finite mixtures to define a fuzzy clustering. We use the ordered stereotype model introduced by Anderson (1984) in order to formulate the ordinal approach, which has rarely been used so far. Two possible reasons for this lack of use might be the absence of standard software for model fitting and its unusual structure including the product of parameters in the linear predictor (Kuss, 2006). The plan of the article is as follows. Section 2 has definitions of the models and its formulation including fuzzy clustering via finite mixtures. Model fitting by using the iterative EM algorithm is described in Section 3. Section 4 presents a review of several model comparison measures and a comparison of eleven information criteria performance. Two real-life examples and simulation studies are given in Section 5, and we conclude with a discussion in Section 6.

2. Model formulation

In this section, we give the standard definition of the ordered stereotype model (Section 2.1) followed by a modification to include clustering (Section 2.2). The likelihood for the suite of basic models is provided next (Section 2.3).

2.1. Data and ordered stereotype model definition

For a set of m ordinal response variables each with q categories measured on a set of n units, the data can be represented by a $n \times m$ matrix Y where, for instance, the n rows represent the subjects of the study and the m columns are the different questions in a particular questionnaire. Although the number of categories might be different, we assume the same q for all such questions. If each answer is a selection from q ordered categories (e.g. strongly agree, agree, neutral, disagree, strongly disagree), then

$$y_{ij} \in \{1, \dots, q\}, \quad i = 1, \dots, n, \ j = 1, \dots, m.$$

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004

The ordered stereotype model (Anderson, 1984) for the probability that y_{ij} takes the category k is characterized by the following log odds

$$\log\left(\frac{P\left[y_{ij}=k \mid \boldsymbol{x}\right]}{P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}\right)=\mu_k+\phi_k\boldsymbol{\delta}'\boldsymbol{x},\quad i=1,\dots,n,\ j=1,\dots,m,\ k=2,\dots,q, \tag{1}$$

where the inclusion of the following monotone increasing constraint

$$0=\phi_1\le\phi_2\le\cdots\le\phi_q=1 \tag{2}$$

preserves the variable response Y is ordinal (see Anderson, 1984). The vector $\boldsymbol{x}$ is a set of predictor variables which can be categorical or continuous, and the vector of parameters $\boldsymbol{\delta}$ represents the effects of $\boldsymbol{x}$ on the log odds of the response variable for the category k relative to the baseline category. The first category is the baseline category, p is the number of covariates, the parameters $\{\mu_2,\dots,\mu_q\}$ are the *cut points*, and $\{\phi_2,\dots,\phi_q\}$ are the parameters which can be interpreted as the "scores" for the categories of the response variable y_{ij}. We restrict $\mu_1=\phi_1=0$ and $\phi_q=1$ to ensure identifiability. With this construction, the category response probabilities in the ordered stereotype model are as follows

$$P\left[y_{ij}=k \mid \boldsymbol{x}\right]=\frac{\exp(\mu_k+\phi_k\boldsymbol{\delta}'\boldsymbol{x})}{\sum_{\ell=1}^{q}\exp(\mu_\ell+\phi_\ell\boldsymbol{\delta}'\boldsymbol{x})}\quad\text{for } k=1,\dots,q, \tag{3}$$

where the probability for the baseline category, as defined in (3), satisfies

$$P\left[y_{ij}=1 \mid \boldsymbol{x}\right]=1-\sum_{\ell=2}^{q}P\left[y_{ij}=\ell \mid \boldsymbol{x}\right]$$

and therefore, since $\mu_1=\phi_1=0$, this probability can be defined as

$$P\left[y_{ij}=1 \mid \boldsymbol{x}\right]=\frac{1}{1+\sum_{\ell=2}^{q}\exp(\mu_\ell+\phi_\ell\boldsymbol{\delta}'\boldsymbol{x})}.$$

Greenland (1994) showed that the stereotype model is appropriate when the progression of the response variable occurs through various stages. Agresti (2010) (see Chapter 4) showed that the stereotype model is equivalent to an ordinal model, such as the proportional odds version of the adjacent-categories logit model, when the scores $\{\phi_k\}$ are a linear function of the different categories of the response variable. An advantage of the stereotype model is that it is more parsimonious than the baseline-category logit model or the multinomial logistic regression model. In addition, the ordered stereotype model is more flexible than the models including the proportional odds structure such as the version for the cumulative logit model (Agresti, 2010, Section 4.3.4) as a result of the $\{\phi_k\}$ parameters. However, the parameters are more difficult to estimate due to the intrinsic nonlinearity which arises from the product of parameters $\phi_k\boldsymbol{\delta}'\boldsymbol{x}$ in the predictor.

2.2. Ordered stereotype model including clustering

The structure of the linear predictor in the ordered stereotype model can include the predictor variables $\boldsymbol{x}$ as numerical covariates, or they may simply be related to the effect of the row and column on the observation y_{ij}. We consider this latter situation and build up $\boldsymbol{\delta}'\boldsymbol{x}$ only taking into account the row and column effects by using a linear formulation. To do this, we define $\{\alpha_1,\dots,\alpha_n\}$ and $\{\beta_1,\dots,\beta_m\}$ as the sets of parameters quantifying the main effects of the n rows and m columns respectively, and the set $\{\gamma_{11},\dots,\gamma_{nm}\}$ are the associations between the different rows and columns. In this way, we can formulate the following saturated model

$$\log\left(\frac{P\left[y_{ij}=k\right]}{P\left[y_{ij}=1\right]}\right)=\mu_k+\phi_k(\alpha_i+\beta_j+\gamma_{ij}),\quad k=2,\dots,q,\ i=1,\dots,n,\ j=1,\dots,m, \tag{4}$$

where $\sum_{i=1}^{n}\alpha_i=\sum_{j=1}^{m}\beta_j=0$ and we impose sum-to-zero constraints on each row and column of the association (or pattern detection) matrix γ. This model has $2q+nm-4$ independent parameters. The relationship between models (3) and (4) is shown in Appendix A. The most common submodels to formulate from the saturated model are the main effect model ($\gamma_{ij}=0$, with $2q+n+m-5$ parameters), the row effect model ($\beta_j=\gamma_{ij}=0$, $2q+n-4$ parameters), the column effect model ($\alpha_i=\gamma_{ij}=0$, $2q+m-4$ parameters) and the null model ($\alpha_i=\beta_j=\gamma_{ij}=0$, $2q-3$ parameters).

The main problem with the model in (4) is that the specific row and column effects in this suite of models overparametrizes the data structure. This model is not parsimonious and it requires a lot of parameters for describing all the effects. A way to reduce the dimensionality of the problem is to introduce fuzzy clustering via finite mixtures. Hence, we obtain the following model formulation including row clustering, column clustering or biclustering.

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004

- Row clustering

$$\log\left(\frac{P\left[y_{ij}=k \mid i\in r\right]}{P\left[y_{ij}=1 \mid i\in r\right]}\right)=\mu_k+\phi_k(\alpha_r+\beta_j+\gamma_{rj}),\quad k=2,\dots,q,\ r=1,\dots,R,\ j=1,\dots,m.$$

- Column clustering

$$\log\left(\frac{P\left[y_{ij}=k \mid j\in c\right]}{P\left[y_{ij}=1 \mid j\in c\right]}\right)=\mu_k+\phi_k(\alpha_i+\beta_c+\gamma_{ic}),\quad k=2,\dots,q,\ i=1,\dots,n,\ c=1,\dots,C.$$

- Biclustering

$$\log\left(\frac{P\left[y_{ij}=k \mid i\in r, j\in c\right]}{P\left[y_{ij}=1 \mid i\in r, j\in c\right]}\right)=\mu_k+\phi_k(\alpha_r+\beta_c+\gamma_{rc}),\quad k=2,\dots,q,\ r=1,\dots,R,\ c=1,\dots,C,$$

where $R\le n$ is the number of row groups, $C\le m$ the number of column groups, $i\in r$ means row i is classified in the row cluster r and $j\in c$ means column j is classified in the column cluster c. It is important to note that the actual membership of the rows among the R row-clusters and the columns among the C column-clusters is unknown and, therefore, it is considered as missing information. Choosing $R\ll n$ ($C\ll m$) ensures that the number of independent parameters in this model is less than nm. The parameters γ_{rj}, γ_{ic} and γ_{rc} may not be necessary in some models, i.e. models without the interaction between row and column groups, where all rows show similar response patterns over the columns, and vice versa. Further, we define $\{\pi_1,\dots,\pi_R\}$ and $\{\kappa_1,\dots,\kappa_C\}$ as the (unknown) proportions of rows and columns in each row and column group respectively, with $\sum_{r=1}^{R}\pi_r=\sum_{c=1}^{C}\kappa_c=1$. We can view π_r and κ_c as the *a priori* row and column membership probabilities. For the case of the ordered stereotype model including fuzzy biclustering, the model is defined with $(q-1)$ cut point parameters μ_k, $(q-2)$ score parameters ϕ_k, $(R-1)$ row effect parameters α_r, $(C-1)$ column effect parameters β_c, $(R-1)(C-1)$ associations between row and column parameters γ_{rc}, $(R-1)$ row cluster membership parameters π_r and $(C-1)$ column cluster membership parameters κ_c. In that way, we may deduce that the model including fuzzy row clustering has $2q+Rm+(R-1)-4$ independent parameters, the column clustering version has $2q+nC+(C-1)-4$ independent parameters and the biclustering one has $2q+RC+(R-1)+(C-1)-4$ independent parameters.

Finally, in the same way as before, we can formulate the probability of the data response y_{ij} being equal to the category k conditional on the appropriate clustering as,

- Row clustering

$$\theta_{r_ijk}=P\left[y_{ij}=k \mid i\in r\right]=\frac{\exp(\mu_k+\phi_k(\alpha_r+\beta_j+\gamma_{rj}))}{\sum\limits_{\ell=1}^{q}\exp(\mu_\ell+\phi_\ell(\alpha_r+\beta_j+\gamma_{rj}))},\quad k=1,\dots,q,\ r=1,\dots,R,\ j=1,\dots,m. \tag{5}$$

- Column clustering

$$\theta_{ic_jk}=P\left[y_{ij}=k \mid j\in c\right]=\frac{\exp(\mu_k+\phi_k(\alpha_i+\beta_c+\gamma_{ic}))}{\sum\limits_{\ell=1}^{q}\exp(\mu_\ell+\phi_\ell(\alpha_i+\beta_c+\gamma_{ic}))},\quad k=1,\dots,q,\ c=1,\dots,C,\ i=1,\dots,n. \tag{6}$$

- Biclustering

$$\theta_{r_ic_jk}=P\left[y_{ij}=k \mid i\in r, j\in c\right]=\frac{\exp(\mu_k+\phi_k(\alpha_r+\beta_c+\gamma_{rc}))}{\sum\limits_{\ell=1}^{q}\exp(\mu_\ell+\phi_\ell(\alpha_r+\beta_c+\gamma_{rc}))},$$
$$k=1,\dots,q,\ r=1,\dots,R,\ c=1,\dots,C. \tag{7}$$

The inclusion of the interaction term allows for different slopes and possible crossings. The additive version of these models omits the interaction term.

2.3. Basic models. Likelihoods

In this section, we summarize the likelihood functions for the cases of row clustering, column clustering and biclustering. The formulation of the complete data log-likelihood is given in each case.

2.3.1. Row clustering

As we noted in the previous section, the unknown data in the case of the row-clustered model is the actual membership of the rows among the R row-clusters. Thus, the incomplete data likelihood only sums over all possible partitions of rows into R clusters:

$$L(\Omega \mid \{y_{ij}\}) = \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \cdots \pi_{r_n} \prod_{i=1}^{n} \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\theta_{r_ijk}\right)^{I(y_{ij}=k)},$$

where Ω is the parameter vector for the case of row clustering, π_{r_i} is the *a priori* row membership probability of row i, θ_{r_ijk} is the probability of the data response defined in (5). Assuming independence among rows and, conditional on the rows, independence over the columns, we can simplify the previous incomplete data likelihood to

$$L(\Omega \mid \{y_{ij}\}) = \prod_{i=1}^{n} \left[\sum_{r=1}^{R} \pi_r \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\theta_{rjk}\right)^{I(y_{ij}=k)} \right].$$

We define the unknown row group memberships through the following indicator latent variables,

$$Z_{ir} = I(i \in r) = \begin{cases} 1 & \text{if } i \in r \\ 0 & \text{if } i \notin r \end{cases} \quad i = 1, \dots, n, \ r = 1, \dots, R, \tag{8}$$

where $i \in r$ indicates that row i is in row group r. It follows that

$$\sum_{r=1}^{R} Z_{ir} = 1, \quad i = 1, \dots, n,$$

and since their *a priori* row membership probabilities are $\{\pi_r\}$

$$(Z_{i1}, \dots, Z_{iR}) \sim \text{Multinomial}(1; \pi_1, \dots, \pi_R), \quad i = 1, \dots, n.$$

These indicator latent variables fulfill the following convenient identity

$$\prod_{r=1}^{R} a_i^{Z_{ir}} = \sum_{r=1}^{R} a_i Z_{ir} \quad \text{for any } a_i \neq 0.$$

Consequently, the complete data log-likelihood of this model using the known data $\{y_{ij}\}$ and the unknown data $\{z_{ir}\}$ is as follows

$$l_c(\Omega \mid \{y_{ij}\}, \{z_{ir}\}) = \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} \log(\pi_r) + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{r=1}^{R} z_{ir} I(y_{ij} = k) \log\left(\theta_{rjk}\right). \tag{9}$$

2.3.2. Column clustering

The model for the case of clustering the columns but not the rows is similar. It assumes independence among columns and, conditional on the columns, independence over the rows. Analogous to Z_{ir} for row clustering (see (8)) we define the following indicator latent variables for the unknown data

$$X_{jc} = I(j \in c) = \begin{cases} 1 & \text{if } j \in c \\ 0 & \text{if } j \notin c \end{cases} \quad j = 1, \dots, m, \ c = 1, \dots, C. \tag{10}$$

The complete data log-likelihood of this model using the known data $\{y_{ij}\}$ and the unknown data $\{x_{jc}\}$ is as follows

$$l_c(\Omega \mid \{y_{ij}\}, \{x_{jc}\}) = \sum_{j=1}^{m} \sum_{c=1}^{C} x_{jc} \log(\kappa_c) + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{c=1}^{C} x_{jc} I(y_{ij} = k) \log\left(\theta_{ick}\right), \tag{11}$$

where Ω is the parameter vector for the case of column clustering and κ_c is the *a priori* column membership probability.

2.3.3. Biclustering

In the case of clustering the rows and the columns simultaneously, the incomplete data likelihood sums over all possible partitions of rows into R clusters and over all possible partitions of columns into C clusters, and is given by

$$L(\Omega \mid \{y_{ij}\}) = \sum_{c_1=1}^{C} \cdots \sum_{c_m=1}^{C} \kappa_{c_1} \cdots \kappa_{c_m} \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \cdots \pi_{r_n} \prod_{i=1}^{n} \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\theta_{r_ic_jk}\right)^{I(y_{ij}=k)}.$$

Here Ω is the parameter vector for the case of biclustering and $\theta_{r_ic_jk}$ is the probability of the data response expressed in (7). Assuming independence among rows and, conditional on the rows, independence over the columns, we can simplify the previous incomplete data likelihood to

$$L(\Omega \mid \{y_{ij}\}) = \sum_{c_1=1}^{C} \cdots \sum_{c_m=1}^{C} \kappa_{c_1} \cdots \kappa_{c_m} \prod_{i=1}^{n} \left[\sum_{r=1}^{R} \pi_r \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\theta_{rjk}\right)^{I(y_{ij}=k)} \right], \tag{12}$$

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004

which sums over the possible column cluster partitions. Similarly, if we assume independence among columns and, conditional on the columns, independence over the rows, we obtain the following simplified expression:

$$L(\Omega \mid \{y_{ij}\}) = \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \cdots \pi_{r_n} \prod_{j=1}^{m} \left[\sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} (\theta_{ick})^{I(y_{ij}=k)} \right]. \tag{13}$$

We define the unknown data through the indicator latent variables described in (8) and (10). Consequently, the complete data log-likelihood of this model using the known data $\{y_{ij}\}$ and the unknown data $\{z_{ir}\}$ and $\{x_{jc}\}$ is as follows:

$$\begin{aligned} l_c(\Omega \mid \{y_{ij}\}, \{z_{ir}\}, \{x_{jc}\}) &= \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{r=1}^{R} \sum_{c=1}^{C} z_{ir} x_{jc} I(y_{ij} = k) \log(\theta_{rck}) \\ &\quad + \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} \log(\pi_r) + \sum_{j=1}^{m} \sum_{c=1}^{C} x_{jc} \log(\kappa_c). \end{aligned} \tag{14}$$

We estimate the MLEs from this expression by using the EM algorithm. In the E-step, the expected value of the first term is approximated using the variational approximation employed by Govaert and Nadif (2005) (see Appendix C for details). With the aim of ensuring a solution avoiding approximations, we use the resulting MLEs from the EM algorithm as starting points to numerically maximize the incomplete-data log-likelihood (12) (or (13)). We note that during the maximization a convenient transformation for the row and column membership parameters $\{\pi_r\}$ and $\{\kappa_c\}$ is $s_r = \text{logit}(\pi_r / \sum_{\ell=r}^{R} \pi_\ell)$ for $r = 1, \ldots, R-1$ and $q_c = \text{logit}(\kappa_c / \sum_{\ell=c}^{C} \kappa_\ell)$ for $c = 1, \ldots, C-1$ respectively. This transformation means that the parameters s_r and q_c are unconstrained, taking values over the whole real line.

3. Estimation of the parameters

In this section, we develop a model fitting procedure using the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997). One of the most common uses of the EM algorithm is in the case of the estimation of the parameters for a finite mixture-density model with incomplete data which in this case is the actual unknown cluster membership of each row and/or column. This method performs a fuzzy assignment of rows and/or columns to clusters based on the posterior probabilities. In this section, we develop this in detail for the case of clustering the rows but not the columns. It has a easy interpretation which helps explain of our methodology. The development for other two cases: clustering the columns but not the rows and biclustering are described in the Appendices B and C.

3.1. The expectation step (E-Step). Row clustering

We apply the E-Step in the EM algorithm by considering the Z_{ir} as latent variables. In this manner, we use their *a priori* probabilities $\{\pi_r\}$ and the current values for the parameters so as to evaluate their expected values, $\widehat{Z_{ir}}$, which are the posterior probabilities that row i is a member of row group r. The conditional expectation of the complete data log-likelihood at iteration t can be expressed as follows

$$\begin{aligned} Q(\Omega \mid \Omega^{(t-1)}) &= E_{\{z_{ir}\}|\{y_{ij}\},\Omega^{(t-1)}} \left[\ell_c(\Omega \mid \{y_{ij}\}, \{z_{ir}\}) \right] \\ &= \sum_{i=1}^{n} \sum_{r=1}^{R} \log(\pi_r^{(t-1)}) E\left[z_{ir} \mid \{y_{ij}\}, \Omega^{(t-1)} \right] \\ &\quad + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{r=1}^{R} I(y_{ij} = k) \log\left(\theta_{rjk}^{(t-1)} \right) E\left[z_{ir} \mid \{y_{ij}\}, \Omega^{(t-1)} \right]. \end{aligned} \tag{15}$$

The latent variable Z_{ir} is a Bernoulli random variable so that

$$E\left[z_{ir} \mid \{y_{ij}\}, \Omega^{(t-1)} \right] = P\left[z_{ir} = 1 \mid \{y_{ij}\}, \Omega^{(t-1)} \right],$$

and applying Bayes' rule to this expression we obtain

$$\begin{aligned} \widehat{Z}_{ir}^{(t)} = P\left[z_{ir} = 1 \mid \{y_{ij}\}, \Omega^{(t-1)} \right] &= \frac{P\left(\{y_{ij}\}, \Omega^{(t-1)} \mid z_{ir} = 1 \right) P(z_{ir} = 1)}{\sum_{\ell=1}^{R} P\left(\{y_{ij}\}, \Omega^{(t-1)} \mid z_{i\ell} = 1 \right) P(z_{i\ell} = 1)} \\ &= \frac{\widehat{\pi}_r^{(t-1)} \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\widehat{\theta}_{rjk}^{(t-1)} \right)^{I(y_{ij}=k)}}{\sum_{\ell=1}^{R} \left\{ \widehat{\pi}_\ell^{(t-1)} \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\widehat{\theta}_{\ell jk}^{(t-1)} \right)^{I(y_{ij}=k)} \right\}}. \end{aligned} \tag{16}$$

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004

This is the expected value of the latent variable Z_{ir} which defines the posterior probability that row i is in group r once we have observed $\{y_{ij}\}$. Finally, we complete the E-step by substituting the previous expression in the complete data log-likelihood at the iteration t expressed in (15),

$$\widehat{Q}(\Omega \mid \Omega^{(t-1)}) = \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{Z}_{ir}^{(t)}\log(\widehat{\pi}_r^{(t-1)}) + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{r=1}^{R}\widehat{Z}_{ir}^{(t)} I(y_{ij}=k)\log\left(\widehat{\theta}_{rjk}^{(t-1)}\right). \tag{17}$$

3.2. The maximization step (M-step). Row clustering

The M-step of the EM algorithm is the global maximization of the previous expression (17) obtained in the E-step. For the case of finite mixture models, the updated estimation of the term containing the row-cluster proportions $\{\pi_1, \dots \pi_R\}$ and the one containing the rest of the parameters Ω are computed independently. Thus, the M-step has two separate parts.

Finally, the maximum-likelihood estimator for the parameter π_r in the case that the indicator variables $\{Z_{1r}, \dots, Z_{nr}\}$ were observable is

$$\widehat{\pi}_r = \frac{1}{n}\sum_{i=1}^{n} z_{ir}, \quad r = 1, \dots, R.$$

However, the data z_{ir} are unobserved in our case. In that manner, we use their conditional expectation which we found in the E-step (16) to replace in the previous expression for the iteration t,

$$\widehat{\pi}_r^{(t)} = \frac{1}{n}\sum_{i=1}^{n} E\left[z_{ir} \mid \{y_{ij}\}, \Omega^{(t-1)}\right] = \frac{1}{n}\sum_{i=1}^{n}\widehat{Z}_{ir}^{(t)}, \quad r = 1, \dots, R. \tag{18}$$

To estimate the remaining parameters Ω, we must numerically maximize the conditional expectation of the complete data log-likelihood (15). In the case of row clustering,

$$\widehat{\Omega} = \underset{\Omega}{\operatorname{argmax}}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{r=1}^{R}\widehat{Z}_{ir} I(y_{ij}=k)\log\left(\theta_{rjk}\right)\right],$$

where the maximization is conditional on the constraints on the parameters. We repeat the two step iteration of the EM algorithm until convergence, that is until there is a small relative change in the likelihood between two consecutive iterations:

$$\frac{\|L(\Omega^{(t+1)} \mid \{y_{ij}\}) - L(\Omega^{(t)} \mid \{y_{ij}\})\|}{\|L(\Omega^{(t)} \mid \{y_{ij}\})\|} \approx 0.$$

A disadvantage of mixture modeling is that the associated likelihood surface may be multimodal. A comprehensive search over different starting points is used to avoid finding only a local maximum. Particularly in our case, the iterative process is repeated 10 times with random starting points and the best MLE (those that lead to higher log-likelihood value) are kept. We have run experiments testing up to 100 random starting points and it was sufficient with 10 repetitions to avoid convergence to local optima.

Finally, we have implemented the EM algorithm for the ordered stereotype model including clustering via finite mixtures and set up the simulation study by using the statistical package **R** 2.15.1 (R Development Core Team, 2010). The maximization was carried out by using the quasi-Newton method provided as an option in `optim()`.

3.3. Reparametrization of the score parameters

The increasing constraint that enforces the scores $\phi_1, \dots, \phi_q$ to be increasing in the stereotype model defined in (2) must be imposed during the estimation procedure. Such a constraint is complex to impose during optimization and hence for convenience we reparametrize $\phi_1, \dots, \phi_q$ as follows.

We first set $\nu_k = \operatorname{logit}(\phi_k)$ for $k = 2, \dots, q-1$, which implies that

$$-\infty \le \nu_2 \le \nu_3 \le \cdots \le \nu_{q-1} \le \infty.$$

We then set

$$\nu_k = \nu_{k-1} + e^{u_k} \quad \text{for } -\infty < u_k < \infty,\ k = 3, \dots, q-1.$$

In that manner, our parameter vector $\{\phi_1 = 0, \phi_2, \dots, \phi_{q-1}, \phi_q = 1\}$ is replaced with $\{\nu_2, u_3, \dots, u_{q-1}\}$ which has the same number of parameters but it is more convenient because the new parameter vector $\mathbb{R}^{q-2}$ is completely unconstrained. This makes the optimization process more straightforward. Once we find the MLEs of $\nu_2, u_3, \dots, u_{q-1}$, we can transform back to the original set of parameters by

$$\phi_k = \begin{cases} 0 & k = 1 \\ \dfrac{1}{1+e^{-\nu_2}} & k = 2 \\ \operatorname{expit}\left[\operatorname{logit}(\phi_2) + \displaystyle\sum_{\ell=3}^{k} e^{u_\ell}\right] & k = 3, \ldots, q-1 \\ 1 & k = q, \end{cases}$$

where $\operatorname{expit}(x) = (1+e^{-x})^{-1}$ is the inverse of the logit function.

4. Model comparison

4.1. Introduction

There are two main approaches to the comparison of a set of candidate likelihood-based models once they are fitted, in order to decide which one (or group of them) best approximates the (unknown) true model. One approach is to carry out a hypothesis test by using the likelihood ratio as a test statistic (LRT). Another approach uses information criteria which are based on a penalized form of the likelihood function where the penalty increases as the number of parameters in the model increases. Some of the most common information criteria measures are Akaike's Information Criterion (AIC, Akaike, 1973), its small-sample modification (AIC$_c$, Akaike, 1973; Hurvich and Tsai, 1989; Burnham and Anderson, 2002), the Bayes' Information Criterion (BIC, Schwarz, 1978) or its Integrated Classification Likelihood version (ICL-BIC, Biernacki et al., 1998).

Unlike the LRT, information criteria quantify the differences in goodness of fit between a set of candidate likelihood-based models (comparative measure of fit), but give no absolute measure of fit. The use of LRT is more computationally demanding than the information criteria because the LRT requires bootstrapping to obtain the p-value. This quantification of the significance is not assessed with the information criteria. However, the LRT does not lead to a suitable significance test in our approach. This occurs because regularity conditions do not hold for $-2\log(\text{LRT})$ to have its usual asymptotic distribution under the null hypothesis for mixture densities. Thus, model selection by LRT tends to overestimate the number of clusters (Stahl and Sallis, 2012). There has been a lot of published research to formulate theoretical results on the null distribution of the LRT for finite mixture model through simulation and bootstrapping studies (see the review in McLachlan and Peel, 2000, Section 6.5). One of the most common way may be using randomization tests (McLachlan, 1987; Manly, 2007; Gotelli and Graves, 1996) to obtain the asymptotic null distribution. However, there is a lack of research on this topic focused on mixtures based on densities from ordinal variables and it might be a field to explore for future research.

4.2. Simulation study

We set up a simulation study to empirically establish a relationship between our likelihood-based methodology for ordinal data and the performance of eleven information criteria in order to determine which was most reliable. We evaluate the following information criteria's performances: AIC, AIC$_c$, BIC, ICL-BIC, AIC$_u$ (McQuarrie et al., 1997), AIC3 (Bozdogan, 1994), CLC (Biernacki and Govaert, 1997), CAIC (Bozdogan, 1987), NEC (Biernacki et al., 1999), AWE (Banfield and Raftery, 1993) and $\mathcal{L}$ (Figueredo and Jain, 2002). Their definitions are given in Table 1.

The results we are interested in are the percentage of simulated experiments where the eleven information criteria correctly determine the true number of row/column clusters in a set of diverse scenarios. The scenarios are determined by varying the sample size/subjects ($n = 50, 100, 500$) and number of measures/questions ($m = 5, 10$). In addition, we made variations in the number of row clusters ($R = 2, 3, 4$), column clusters ($C = 2, 3, 4$) and the space between the $q = 4$ score parameters $\{\phi_k\}$. The five scenarios for $\{\phi_k\}$ may be described by: equal spacing between any pair of adjacent score parameters (Scenario 1), one pair of adjacent score parameters are very close in value (Scenario 2), one of the mixing cluster proportions is close to zero (Scenario 3), one pair of adjacent score parameters have the same value (Scenario 4), and the same as the first scenario but increasing the number of measures to $m = 10$ (Scenario 5). All the parameters for each scenario in the row clustering and biclustering cases are shown in Tables D.10 and D.11 in Appendix D.

For each scenario, we drew ($h = 100$) data sets, and selected the best model for each data set using each information criterion. Therefore, we worked with a total of 4500 and 6000 data sets for row clustering and biclustering respectively. The EM algorithm to obtain the estimators is repeated 10 times with random starting points and the estimates with the highest likelihood are kept.

Fig. 1 is a histogram displaying the percentage of cases in which each information criterion determines the true number of row clusters across the five scenarios and the factors used in the experimental control. The best performance was AIC (correctly selecting the number of row clusters in 93.8% of cases), followed by AIC$_c$ (89.8%) and AIC$_\text{u}$ (82.4%). In the case of biclustering, the results are very similar as AIC also performs the best, although with a lower percentage of correctly selecting the number of row and column clusters than the row clustering case (86.1%). AIC$_c$ and AIC$_\text{u}$ also perform very well with percentages close to AIC: 85.6% and 84.2% respectively. BIC is underestimating the number of clusters (incorrectly

Table 1
Information criteria summary table for one-dimension clustering case.

Criteria	Definition	Proposed for	Depending on
AIC (Akaike, 1973)	$-2\ell + 2K$	Regression	nm
AIC$_c$ (Akaike, 1973)	$AIC + \frac{2K(K+1)}{nm-K-1}$		K and nm
AIC$_u$ (McQuarrie et al., 1997)	$AIC_c + nm\log\left(\frac{nm}{nm-K-1}\right)$		
CAIC (Bozdogan, 1987)	$-2\ell + K(1+\log(nm))$		
BIC (Schwarz, 1978)	$-2\ell + K\log(nm)$		
AIC3 (Bozdogan, 1994)	$-2\ell + 3K$	Clustering	K
CLC (Biernacki and Govaert, 1997)	$-2\ell + 2EN(R)$		$EN(\cdot)$
NEC(R) (Biernacki et al., 1999)	$\frac{EN(R)}{\ell(R)-\ell(1)}$		
ICL-BIC (Biernacki et al., 1998)	$-2\ell_c + K\log(nm)$		K, nm and $EN(\cdot)$
AWE (Banfield and Raftery, 1993)	$-2\ell_c + 2K\left(\frac{3}{2} + \log(nm)\right)$		
$\mathcal{L}$ (Figueredo and Jain, 2002)	$-\ell - \frac{K}{2}\sum\log(\frac{nm\pi_R}{12}) - \frac{R}{2\log(\frac{nm}{12})} - \frac{R(K+1)}{2}$		K, nm and π_R

Notes: nm is the total sample size which is the number of elements in the response matrix Y. K is the number of parameters, R the number of clusters, π_R the mixing cluster proportion, ℓ the maximized incomplete data log-likelihood, ℓ_c is the maximized complete data log-likelihood (see Eq. (9) for row clustering and Eq. (11) for column clustering). $EN(\cdot)$ is the entropy function defined by $EN(R) = \ell - \ell_c$.

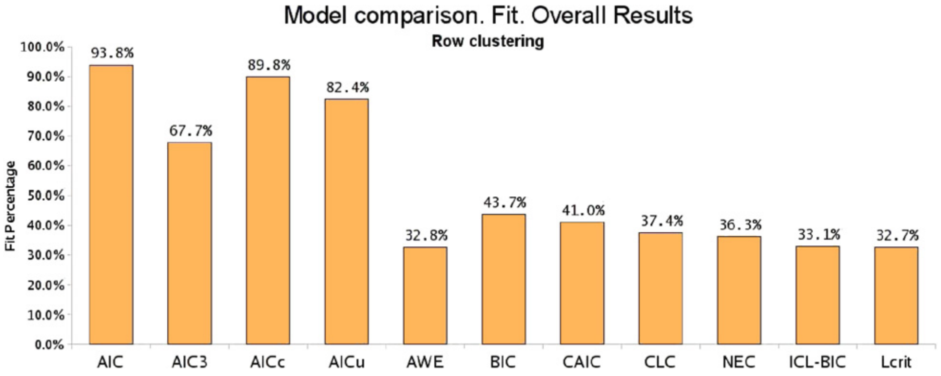


Fig. 1. Simulation study results for row clustering. Bars depict the percentage of cases for each information criterion fits the true number of row clusters.

selecting a smaller number of clusters in 56% and 63.2% of cases in row clustering and biclustering respectively). A very poor performance is obtained by ICL-BIC (correctly selecting the number of clusters in 33.1% and 31.3% of cases in row clustering and biclustering). Our results are in accordance with Fonseca and Cardoso (2007) for the categorical case.

Tables 2 and 3 show the best 5 information criteria performances in the case of row clustering and biclustering respectively. In both cases, we can observe that AIC is the best measure over the 5 scenarios and the ranking positions are exactly the same over the 5 scenarios. The best performance is scenario 5 which has the largest number of measures m. On the other hand, the worse achievement is scenario 3 which has one of the mixing cluster proportion close to zero and, therefore, the percentage of underestimated number of clusters is higher. Regardless of this challenging scenario configuration, the AIC and AIC$_c$ performances are still quite satisfactory (over 75% of fitted cases in row and biclustering).

Our conclusion is that AIC is the best information criteria when dealing with ordinal data and we fit likelihood-based finite mixture models with the ordinal stereotype model as the components in the mixture. It is important to remark these results are just evaluating the fact of obtaining the right number of clusters in the mixture, but it does not imply that they are the best clustering structure for the data.

5. Results

In this section, the reliability of estimation of the stereotype model parameters is demonstrated in a simulation study (Section 5.1). In addition, we illustrate the stereotype model and our likelihood-based clustering method with two real-

Table 2
Model comparison simulation study results. Row clustering. Ranking of the best 5 information criteria measures.

	Overall	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
AIC	**93.8**%	91.4%	97.6%	88.0%	92.9%	99.1%
AIC$_c$	**89.8**%	90.2%	94.8%	74.7%	91.1%	98.2%
AIC$_u$	**82.4**%	79.0%	80.0%	66.7%	88.0%	98.2%
AIC3	**67.7**%	61.7%	65.6%	56.7%	56.4%	98.2%
BIC	**43.7**%	41.2%	39.1%	40.0%	39.6%	58.7%

Table 3
Model comparison simulation study results. Biclustering. Ranking of the best 5 information criteria measures.

	Overall	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
AIC	**86.1**%	89.2%	82.3%	80.5%	85.5%	92.8%
AIC$_c$	**85.6**%	89.2%	81.5%	80.0%	84.5%	92.8%
AIC$_u$	**84.2**%	84.8%	80.7%	79.3%	83.3%	92.8%
AIC3	**71.2**%	75.8%	65.5%	64.7%	66.5%	83.3%
BIC	**36.5**%	34.5%	35.2%	33.5%	32.3%	47.2%

life examples (Section 5.2). In order to do model selection, the two best information criteria (AIC and AIC$_c$) according to the comparison study (Section 4) are computed together with the most commonly used BIC and ICL-BIC. Thus, their performances can be compared.

5.1. Simulation study

We set up a simulation study to test how reliably, in a diverse range of scenarios, we were able to estimate the parameters of stereotype models using the EM algorithm. We are not testing model selection here (that was tested in Section 4): we simulate data sets and then fit the correct model to those data.

The simulation program was written in **R**, wrapping the likelihood function which was written in **C**. The design of the study refers to an ordinal response variable with four categories and we varied the sample size ($n = 25, 50, 200, 500, 1000, 5000$), the number of columns ($m = 5, 10, 15$) and the number of row and/or column clusters ($R = 2, 3, 4, 5, 6$ and $C = 2, 3$). For each combination of sample size and number of row clusters, a single set of parameters values was chosen and 100 data sets (replicates) were generated. The MLEs and their standard errors were found for each replicate. The general results for the score parameters $\{\widehat{\phi_k}\}$ for row clustering without interaction factors are given in Table 4. Tables 5 and 6 present the equivalent results for column clustering and biclustering respectively. The simulation scenarios including the interaction factors for row clustering and biclustering version are shown in Tables E.12 and E.13 in Appendix E. In each case the tables show the mean of the MLEs and of their corresponding standard errors over the 100 replicates.

For all models (row clustering, column clustering and biclustering) the estimates of the parameters $\{\phi_k\}$, $\{\pi_r\}$ and $\{\kappa_c\}$ are close to their true values and as expected the variability decreases with increasing the sample size n, and the number of columns m in the case of column clustering. Fig. 2 shows the 100 separate estimates of $\widehat{\phi_2}$ and $\widehat{\phi_3}$ for the row clustering model with $R = 2$ row clusters plotted against each other for varying sample sizes. Note that all the estimates in the figure show the ordering constraint $\phi_2 < \phi_3$, which restricts the estimates to the upper left triangle of the plot. This sequence of plots shows that the estimation process consistently returned MLEs for the score parameters $\{\phi_k\}$ close to their true values (the diamond point in each plot) with reducing standard error as the sample size increases. Figs. E.11 and E.12 in Appendix E show similar results for the column clustering model with $C = 2$ clusters and the biclustering model with $R = 2$ and $C = 2$ clusters respectively. However, the column clustering model has the drawback that the number of $\{\alpha_i\}$ parameters is large when the sample size n is increased (e.g., 156 parameters with $n = 50$, $q = 4$ and $C = 3$) and therefore estimators would be poor with large sample sizes in that case. The consequences of this are that the standard errors are slightly higher than for row clustering and biclustering even as the number of columns m increases.

In addition, we have observed that the EM algorithm converges to a point far away from the true value. We do not notice this problem in the row clustering and biclustering versions but we detected it in approximately 5% of cases with column clustering when the sample size is $n = 50$. This problem is apparently caused by the large number of individual row parameters $\{\alpha_i\}$ in column clustering. Another inherent drawback in finite mixtures is that the likelihood has a multimodal surface.

Our initial results described above are encouraging in their ability to estimate parameters correctly. However, we were interested to test the success of the estimation in challenging situations where it might be expected that estimation might be difficult. We chose two particular scenarios. The first case is when two of the score parameters $\{\phi_k\}$ have equal values and, therefore, from the point of view of detecting clustering, we could merge their corresponding response categories. A second scenario is to set a very small *a priori* membership probability, e.g. $\pi_2 = 0.015$, and, consequently, few data units will be classified in the related cluster. The chosen probability must not be related to the first or last response categories because there is a relationship with the score parameters (see (18)) and their corresponding score parameters are set to $\phi_1 = 0$ and $\phi_q = 1$ to avoid identifiability problems. Therefore, it is more interesting to test a free score parameter.

Table 4
Simulation study. Estimated score parameters for stereotype model including row clustering without interaction factors $\boldsymbol{\mu_k}+\boldsymbol{\phi_k}(\boldsymbol{\alpha_r}+\boldsymbol{\beta_j})$. MLEs and their standard errors from the score and row membership parameters ($\{\phi_k\}$, $\{\pi_r\}$) for different number of row clusters R and sample sizes n are shown.

R	Numpar	True param.	$n = 200$		$n = 500$		$n = 1000$		$n = 5000$	
			Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
2	11	$\phi_2 = 0.335$	0.366	0.183	0.377	0.114	0.335	0.080	0.336	0.036
		$\phi_3 = 0.672$	0.682	0.188	0.679	0.115	0.670	0.081	0.671	0.036
		$\pi_1 = 0.550$	0.523	0.046	0.541	0.031	0.553	0.019	0.552	0.009
3	13	$\phi_2 = 0.335$	0.330	0.184	0.332	0.114	0.337	0.080	0.335	0.035
		$\phi_3 = 0.672$	0.669	0.169	0.675	0.103	0.673	0.074	0.674	0.032
		$\pi_1 = 0.200$	0.189	0.021	0.194	0.017	0.187	0.010	0.211	0.004
		$\pi_2 = 0.500$	0.529	0.118	0.491	0.121	0.489	0.091	0.496	0.044
4	15	$\phi_2 = 0.335$	0.334	0.160	0.333	0.102	0.331	0.071	0.334	0.032
		$\phi_3 = 0.672$	0.682	0.158	0.670	0.100	0.668	0.069	0.671	0.031
		$\pi_1 = 0.150$	0.261	0.097	0.080	0.037	0.146	0.028	0.151	0.022
		$\pi_2 = 0.300$	0.241	0.131	0.332	0.048	0.288	0.028	0.289	0.016
		$\pi_3 = 0.250$	0.255	0.133	0.290	0.048	0.263	0.015	0.244	0.008
5	17	$\phi_2 = 0.335$	0.331	0.178	0.335	0.110	0.331	0.076	0.336	0.034
		$\phi_3 = 0.672$	0.678	0.180	0.675	0.112	0.671	0.077	0.673	0.034
		$\pi_1 = 0.150$	0.153	0.027	0.146	0.031	0.145	0.015	0.145	0.003
		$\pi_2 = 0.300$	0.313	0.058	0.326	0.049	0.295	0.027	0.288	0.009
		$\pi_3 = 0.100$	0.092	0.026	0.089	0.032	0.094	0.099	0.102	0.003
		$\pi_4 = 0.200$	0.217	0.032	0.205	0.023	0.199	0.014	0.202	0.003
6	19	$\phi_2 = 0.335$	0.325	0.193	0.336	0.121	0.322	0.086	0.333	0.060
		$\phi_3 = 0.672$	0.671	0.194	0.673	0.119	0.656	0.083	0.671	0.059
		$\pi_1 = 0.150$	0.156	0.033	0.150	0.023	0.139	0.007	0.140	0.004
		$\pi_2 = 0.300$	0.296	0.038	0.302	0.035	0.294	0.010	0.290	0.005
		$\pi_3 = 0.100$	0.093	0.039	0.090	0.027	0.095	0.006	0.096	0.004
		$\pi_4 = 0.200$	0.203	0.034	0.204	0.026	0.200	0.004	0.200	0.003
		$\pi_5 = 0.150$	0.158	0.019	0.161	0.015	0.162	0.006	0.160	0.003

Table 5
Simulation study. Estimated score parameters for stereotype model including column clustering without interaction factors $\boldsymbol{\mu_k}+\boldsymbol{\phi_k}(\boldsymbol{\alpha_i}+\boldsymbol{\beta_c})$. MLEs and their standard errors from the score and column membership parameters ($\{\phi_k\}$, $\{\kappa_c\}$) for different number of column clusters C, number of columns m and sample sizes n are shown.

C	Numpar	True param.	$m = 5$		$m = 10$		$m = 15$	
			Mean	S.E.	Mean	S.E.	Mean	S.E.
			$n = 25$					
2	31	$\phi_2 = 0.335$	0.291	0.261	0.314	0.143	0.329	0.100
		$\phi_3 = 0.672$	0.722	0.245	0.652	0.169	0.681	0.103
		$\kappa_1 = 0.600$	0.589	0.190	0.589	0.122	0.588	0.095
3	33	$\phi_2 = 0.335$	0.296	0.259	0.307	0.158	0.342	0.090
		$\phi_3 = 0.672$	0.790	0.283	0.712	0.177	0.682	0.110
		$\kappa_1 = 0.400$	0.371	0.204	0.376	0.124	0.390	0.087
		$\kappa_2 = 0.200$	0.179	0.196	0.195	0.112	0.195	0.086
			$n = 50$					
2	56	$\phi_2 = 0.335$	0.397	0.215	0.348	0.119	0.335	0.081
		$\phi_3 = 0.672$	0.736	0.204	0.704	0.111	0.678	0.075
		$\kappa_1 = 0.600$	0.618	0.176	0.609	0.092	0.599	0.063
3	58	$\phi_2 = 0.335$	0.386	0.211	0.342	0.116	0.332	0.078
		$\phi_3 = 0.672$	0.724	0.227	0.693	0.117	0.675	0.065
		$\kappa_1 = 0.400$	0.377	0.183	0.386	0.086	0.403	0.068
		$\kappa_2 = 0.200$	0.204	0.179	0.201	0.083	0.201	0.055

We have simulated these two specific scenarios for the row clustering, column clustering and biclustering models and Tables E.14–E.16 in Appendix E summarizes the simulation results. These are very satisfactory because our approach can identify these particular scenarios and get back values close to the true score parameters $\{\phi_k\}$ in the suite of models tested. However, some of the approximate 95% confidence intervals for the *a priori* membership probabilities $\{\pi_r\}$ do not cover their true values when the sample size is higher than $n = 1000$ and, therefore, the variability is reduced (e.g., row clustering model with $R = 4$ clusters with statistical theory (central limit theorem) providing an approximate 95% CI for π_3 and $n = 5000$ (Table E.14) is (0.262, 0.298) when the true value is 0.23). In addition, we have observed the same drawbacks described above in the column clustering version.

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004

Table 6
Simulation study. Estimated score parameters for stereotype model including biclustering without interaction factors $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_c})$. MLEs and their standard errors from the score, row and column membership parameters $(\{\phi_k\}, \{\pi_r\}, \{\kappa_c\})$ for different number of row and column clusters R and C and sample sizes n are shown.

R	C	Numpar	True param.	$n = 25$		$n = 50$		$n = 100$	
				Mean	S.E.	Mean	S.E.	Mean	S.E.
2	2	9	$\phi_2 = 0.335$	0.354	0.357	0.351	0.266	0.329	0.142
			$\phi_3 = 0.672$	0.686	0.379	0.658	0.260	0.693	0.143
			$\pi_1 = 0.600$	0.504	0.234	0.671	0.175	0.585	0.139
			$\kappa_1 = 0.400$	0.446	0.231	0.415	0.142	0.409	0.074
2	3	11	$\phi_2 = 0.335$	0.319	0.341	0.322	0.243	0.324	0.132
			$\phi_3 = 0.672$	0.753	0.365	0.693	0.232	0.671	0.142
			$\pi_1 = 0.600$	0.490	0.201	0.522	0.159	0.577	0.086
			$\kappa_1 = 0.400$	0.387	0.209	0.388	0.169	0.411	0.121
			$\kappa_2 = 0.200$	0.229	0.210	0.222	0.177	0.189	0.105
3	2	11	$\phi_2 = 0.335$	0.345	0.337	0.342	0.266	0.334	0.155
			$\phi_3 = 0.672$	0.712	0.302	0.688	0.201	0.669	0.146
			$\pi_1 = 0.300$	0.313	0.209	0.313	0.128	0.301	0.106
			$\pi_2 = 0.400$	0.404	0.200	0.346	0.118	0.367	0.093
			$\kappa_1 = 0.400$	0.381	0.196	0.397	0.131	0.400	0.062
3	3	13	$\phi_2 = 0.335$	0.362	0.341	0.355	0.219	0.337	0.145
			$\phi_3 = 0.672$	0.706	0.300	0.627	0.210	0.664	0.135
			$\pi_1 = 0.300$	0.283	0.202	0.296	0.129	0.311	0.094
			$\pi_2 = 0.400$	0.368	0.181	0.373	0.113	0.398	0.088
			$\kappa_1 = 0.400$	0.388	0.182	0.392	0.095	0.402	0.079
			$\kappa_2 = 0.200$	0.195	0.195	0.197	0.099	0.200	0.081

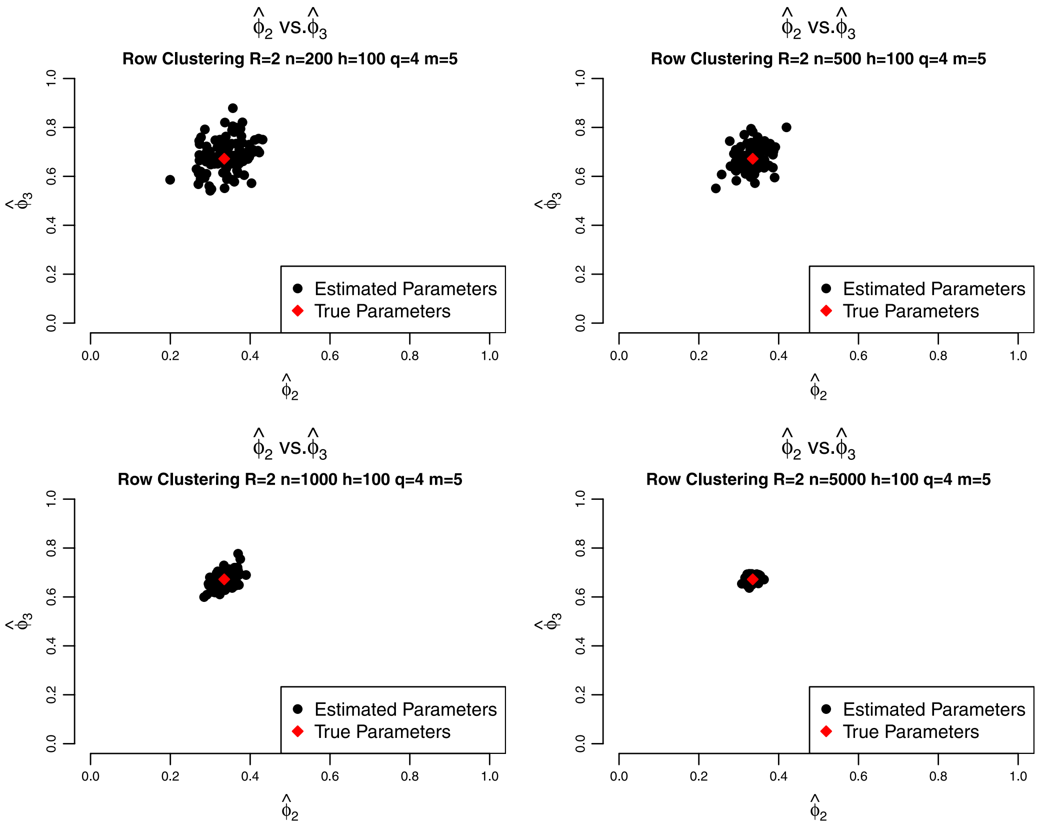
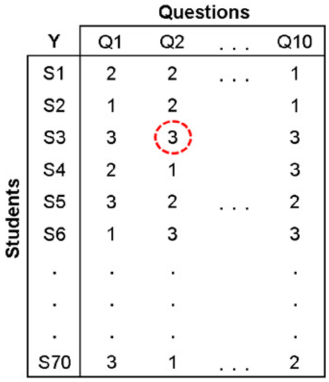


Fig. 2. Simulation study: convergence of $\widehat{\phi_2}$ and $\widehat{\phi_3}$ for the stereotype model including the row clustering $(\alpha_r + \beta_j)$ with $R = 2$ row clusters. n, h, q, m describe the sample size, the number of replicates, the number of categories and the number of covariates respectively. The diamond point represents the true value of the parameter.



		Questions			
	Y	Q1	Q2	. . .	Q10
Students	S1	2	2	. . .	1
	S2	1	2		1
	S3	3	3		3
	S4	2	1		3
	S5	3	2	. . .	2
	S6	1	3		3
	.	.	.		.
	.	.	.		.
	.	.	.		.
	S70	3	1	. . .	2

Fig. 3. Data set with the “Applied Statistics” course feedback forms. The dotted circle indicates the student number 3 answered the question number 2 as “agree” (coded as 3).

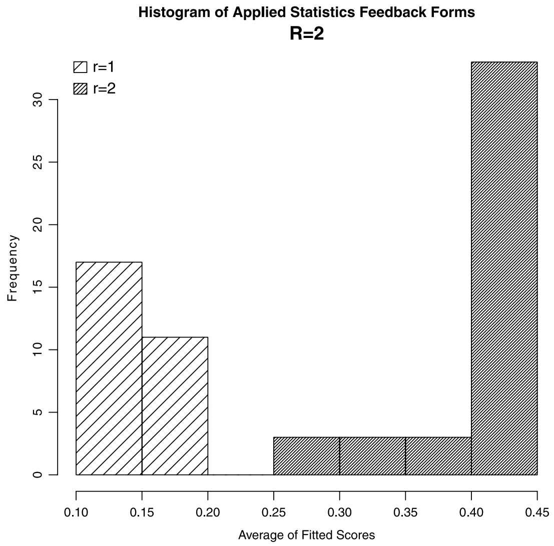


Fig. 4. Histogram of the $R = 2$ fitted student clusters $\{\overline{\phi}_{(i.)}\}$ from the row clustering version model.

5.2. Real-life data examples

5.2.1. Example 1: applied statistics course feedback

The example is corresponding to a data set with the responses of 70 students giving feedback about a second year Applied Statistics course at Victoria University of Wellington. The responses were collected in feedback forms through 10 questions (e.g. “The way this course was organised has helped me to learn”), where each question had three possible ordinal response categories: “disagree” (coded as 1), “neither agree or disagree” (coded as 2) and “agree” (coded as 3). Each question was written so that “agree” indicates a positive view of the course. The whole data set is shown in Table F.18 in Appendix F.

In that way, the dimensions of the data matrix Y with the responses are $n = 70$ rows (students) and $m = 10$ columns (questions) where each observation can take one of the three possible categories. Therefore, we can represent the data in a matrix as shown in Fig. 3.

The main goal is to select the model which best represents the data, including determining the number of different groups in the data. We have fitted a suite of models from the null model (no clustering) to the main effects model and their versions including row clustering, column clustering and biclustering. For each model, the information criteria AIC, AIC$_c$, BIC and ICL-BIC were computed and the results are summarized in Table 7.

AIC and AIC$_c$ indicate that the best models are models with main effects without interaction factors $(\mu_k + \phi_k(\alpha_i + \beta_j))$ with AIC $= 965.26$ and AIC$_c$ $= 987.32$, and the stereotype model version including row clustering with $R = 2$, 3 or 4 row groups and without interaction factors $(\mu_k + \phi_k(\alpha_r + \beta_j))$. Although the main effects model is found to be the best model, for demonstration purposes we discuss here the row clustered models without interaction factors, which have greater interpretability. Figs. 4–6 show three histograms depicting a newly-defined average of the fitted scores of student responses over the 10 questions where each student is allocated to the row group to which she/he belongs with highest

Table 7
Suite of models fitted for applied statistics course feedback forms data set. For each information criterion, the best model in each group (no clustering, row clustering, column clustering and biclustering) is shown in boldface.

Model		R	C	npar	AIC	AIC$_c$	BIC	ICL-BIC
Null model	μ_k	1	1	3	1298.40	1298.46	1312.06	1312.06
Row effects	$\mu_k + \phi_k\alpha_i$	n	1	72	1224.04	1241.30	1551.72	1551.72
Column effects	$\mu_k + \phi_k\beta_j$	1	m	12	1105.50	1106.03	**1160.11**	**1160.11**
Main effects	$\mu_k + \phi_k(\alpha_i + \beta_j)$	n	m	81	**965.26**	**987.32**	1333.90	1333.90
Row clustering	$\mu_k + \phi_k\alpha_r$	2	1	5	1251.70	1251.82	1274.45	1302.34
		3	1	7	1241.60	1241.82	1273.47	1325.84
		4	1	9	1251.56	1251.88	1292.52	1348.77
	$\mu_k + \phi_k(\alpha_r + \beta_j)$	2	m	14	1025.75	1026.45	1089.47	**1109.82**
		3	$\boldsymbol{m}$	**16**	**1013.44**	**1014.33**	**1086.25**	1117.53
		4	m	18	1017.44	1018.56	1099.36	1176.50
	$\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj})$	2	m	23	1042.30	1044.08	1146.98	1167.34
		3	m	34	1032.43	1036.23	1187.17	1219.93
		4	m	45	1020.08	1026.70	1224.88	1244.90
Column clustering	$\mu_k + \phi_k\beta_c$	1	2	5	1279.94	1280.06	**1242.90**	**1302.69**
		1	3	7	**1278.59**	**1278.80**	1310.45	1315.47
	$\mu_k + \phi_k(\alpha_i + \beta_c)$	n	2	74	1409.09	1427.31	1435.93	1745.82
		n	3	76	1430.75	1450.06	1490.43	1776.63
Biclustering	$\mu_k + \phi_k(\alpha_r + \beta_c)$	2	2	7	1115.32	1115.53	1147.18	1182.21
		3	2	9	1110.29	1110.61	1151.25	1192.03
		4	2	11	1114.29	1114.75	1164.36	1206.08
		2	3	9	1060.77	1061.09	**1101.73**	**1138.13**
		3	3	11	**1052.04**	**1052.49**	1102.10	1148.95
		4	3	13	1056.04	1056.65	1115.20	1221.54
		2	4	11	1064.77	1065.23	1114.83	1151.52
		3	4	13	1056.04	1056.65	1115.20	1165.96
		4	4	15	1060.04	1060.84	1128.31	1234.04
	$\mu_k + \phi_k(\alpha_r + \beta_c + \gamma_{rc})$	2	2	8	1117.33	1117.59	1153.73	1188.76
		3	2	11	1098.29	1098.75	1148.35	1204.03
		4	2	14	1104.29	1104.99	1168.01	1278.05
		2	3	11	1064.56	1065.01	1114.62	1151.15
		3	3	15	1058.96	1059.75	1127.22	1184.06
		4	3	19	1127.46	1128.69	1213.93	1325.72
		2	4	14	1070.56	1071.26	1134.28	1174.55
		3	4	19	1066.96	1068.19	1153.43	1214.02
		4	4	24	1076.96	1078.89	1186.18	1285.48

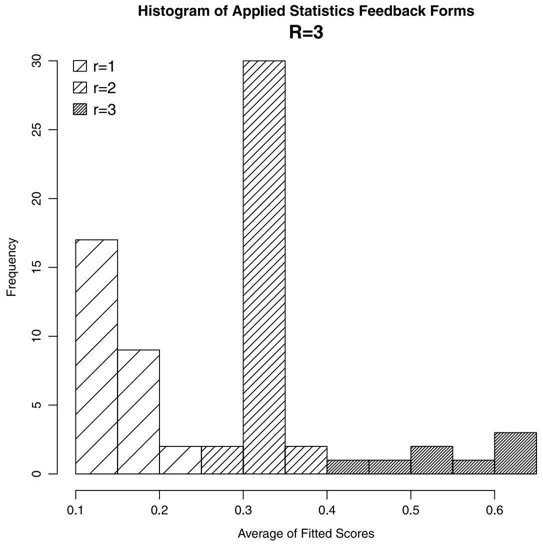


Fig. 5. Histogram of the $R = 3$ fitted student clusters $\{\overline{\phi}_{(i.)}\}$ from the row clustering version model.

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004

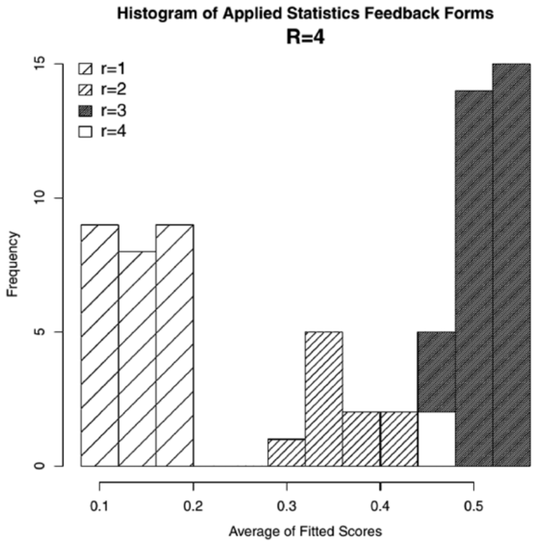


Fig. 6. Histogram of the $R = 4$ fitted student clusters $\{\overline{\phi}_{(i.)}\}$ from the row clustering version model.

posterior probability. Different shade bars represent the row cluster to which the student is assigned according to the corresponding model. This average score (along the x-axis) is calculated in the following way. First, we compute the fitted response probabilities with the estimated parameters over the R row clusters and the q response categories,

$$P[y_{ij} = k \mid i \in r] = \frac{\exp(\widehat{\mu}_k + \widehat{\phi}_k(\widehat{\alpha}_r + \widehat{\beta}_j))}{\sum_{\ell=1}^{q} \exp(\widehat{\mu}_\ell + \widehat{\phi}_\ell(\widehat{\alpha}_r + \widehat{\beta}_j))}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m,\ k = 1, \ldots, q,\ r = 1, \ldots, R.$$

From the previous probabilities, we can compute the weighted average over the q categories for each row cluster

$$\overline{\phi}_{rj} = \sum_{k=1}^{q} \widehat{\phi}_k P[y_{ij} = k \mid i \in r], \quad i = 1, \ldots, n,\ j = 1, \ldots, m,\ r = 1, \ldots, R.$$

From here, we can calculate the mean response level of individual i to question j, conditional on its (fuzzy) allocation to the row clusters:

$$\overline{\phi}_{(ij)} = \sum_{r=1}^{R} \widehat{z}_{ir} \overline{\phi}_{rj}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m. \tag{19}$$

This is a numerical measure of the typical response to question j for members of row group r, appropriately adjusting for the uneven spacing of the levels of the ordinal response. Finally, we determine the mean of the previous weighted averages over the m columns in order to get the average fitted scores of individual i across all of the questions

$$\overline{\phi}_{(i.)} = \frac{1}{m} \sum_{j=1}^{m} \overline{\phi}_{(ij)}, \quad i = 1, \ldots, n.$$

Figs. 4–6 display these $\overline{\phi}_{(i.)}$ values for $R = 2$, 3 and 4 clusters.

Figs. 4 and 5 respectively show two and three clearly distinguished groups. The histogram from Fig. 4 presents two modes and Fig. 5 shows two clear modes and one small mode located in the right-tale. However, Fig. 6 where four groups are fitted shows that the fourth group only includes two students and they are not clearly distinguished from the other three groups. These graphs illustrate the conclusion from AIC/AIC$_c$ that among the row clustering models, the model with three student groups is the best for our data.

Figs. 7 and 8 display the estimated probability $\widehat{\theta}_{rk}$ of a member of group r responding at category level k (Eq. (5)). We might conclude that the students classified in the first group correspond to those with lowest opinion regarding the course, the ones in the second group have a more moderate opinion about the course and the students in the third group are those with more positive (though still heterogeneous) set of opinions.

5.2.2. Example 2: tree presences in Great Smoky Mountains

We use a real data set from community ecology as a second example to illustrate our likelihood-based clustering method. The data set is regarding the distribution of 41 different tree species along 12 different site stations located at altitudes

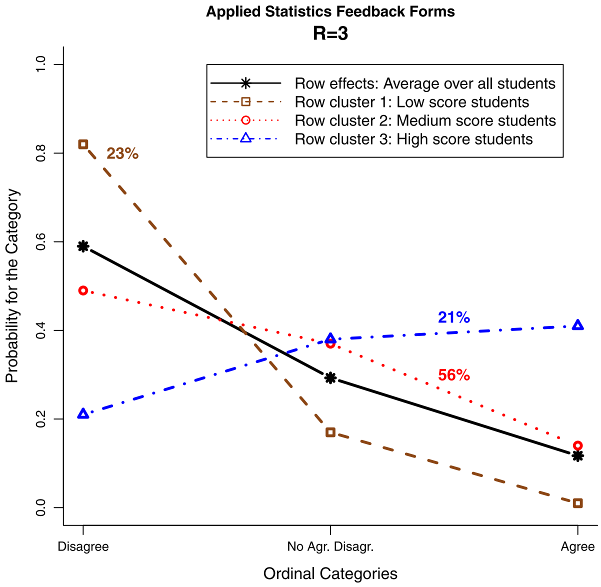


Fig. 7. $R = 3$ student groups. The lines depict the probability for the category $\widehat{\theta}_{r_ijk} = P\left[y_{ij} = k \mid i \in r\right]$ (see Eq. (5)) for each group r and the average over all students (black line). The percentage labeling is the estimated posteriori probability $\widehat{\pi}_r$ that member of each group r will respond to questions in each ordinal category (Eq. (18)).

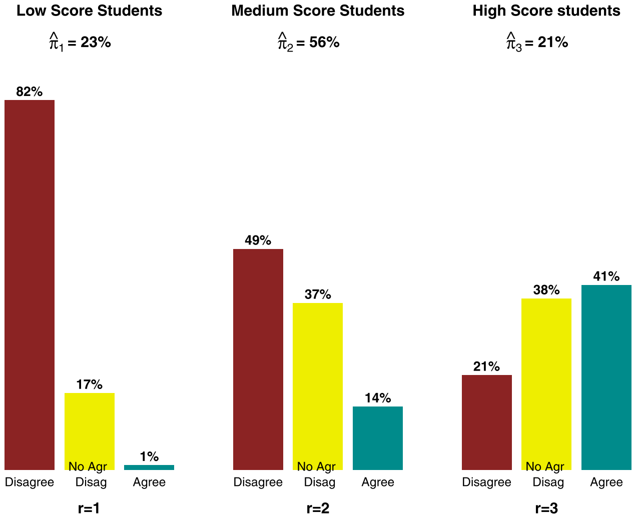


Fig. 8. $R = 3$ student group profiles. The percentage represents the probability $\widehat{\theta}_{r_ijk}$ in each category (Eq. (5)).

between 3500 and 4500 ft and sorted by moisture level (wetter to drier). The observations consist of percentage of total tree species present at each station and was presented in R.H. Whittaker's study of vegetation of the Great Smoky Mountains (Whittaker, 1956, Table 3). The data set is reproduced in Table 8.

The data include cells with a low but nonzero detection, at levels <0.5%. This missing data cause that the effective distances between the values do not reflect their interpretative distance. Thus, transformations may be required (Hennig and Liao, 2013) and that presents an appropriate opportunity to replace numerical data with an ordinal scale. In order to apply our model approach, we transform the original data $\{x_{ij}\}$ regarding tree presence percentage to ordinal response categories setting

$$y_{ij} = \begin{cases} 0 & \text{if } x_{ij} = 0\% \\ 1 & \text{if } 0\% < x_{ij} \leq 0.5\% \\ 2 & \text{if } 0.5\% < x_{ij} \leq 1\% \\ 3 & \text{if } 1\% < x_{ij} \leq 8\% \\ 4 & \text{if } x_{ij} > 8\% \end{cases}$$

Table 8
Great Smoky Mountains data (Whittaker, 1956). Presence distribution of tree species along different site stations. All figures are percentages of total stems presence in station.

Tree species	Station number											
	1	2	3	4	5	6	7	8	9	10	11	12
Fangus grandifolia	10	5	1	1	1	–	–	–	–	–	–	–
Ilex opaca	–	1	–	<0.5	–	–	–	–	–	–	–	–
Picea rubens	–	<0.5	–	<0.5	<0.5	–	–	–	–	–	–	–
Cornus alternifolia	1	1	–	<0.5	<0.5	–	–	–	–	–	–	–
Aesculus octandra	8	9	4	2	6	1	–	–	–	–	–	–
Tilia heterophylla	29	11	9	1	14	3	–	–	–	–	–	–
Acer spicatum	–	16	11	–	17	1	–	–	–	–	–	–
Acer saccharum	17	7	1	1	5	1	–	–	–	–	–	–
Prunus serotina	2	1	–	1	<0.5	2	–	–	–	–	–	–
Fraxinus americana	1	1	–	1	1	<0.5	–	–	–	–	–	–
Betula allegheniensis	5	17	10	15	4	1	<0.5	–	–	–	–	–
Magnolia acuminata	–	<0.5	–	–	<0.5	–	1	–	–	–	–	–
Magnolia fraseri	–	–	20	4	1	–	1	–	–	–	–	–
Tsuga canadensis	20	22	34	62	18	<0.5	<0.5	1	–	–	–	–
Halesia monticola	5	8	4	1	9	13	3	1	1	–	–	–
Ilex montana	1	<0.5	–	1	1	1	2	–	–	–	–	–
Acer pensylvanicum	1	<0.5	1	3	8	3	<0.5	1	–	–	–	–
Amelanchier laevis	–	<0.5	–	<0.5	<0.5	–	–	–	–	–	–	–
Quercus borealis	–	1	–	–	2	40	10	4	15	11	2	1
Acer rubrum	–	1	–	–	1	6	37	21	13	10	8	1
Prunus pensylvanica	–	–	2	–	–	–	1	–	–	–	–	–
Betula lenta	–	–	1	4	4	1	2	2	–	–	–	–
Clethra acuminata	–	–	–	1	<0.5	–	–	–	–	–	–	–
Hamamelis virginiana	–	–	–	–	2	5	17	7	1	–	2	–
Cornus florida	–	–	–	–	1	–	<0.5	4	–	–	–	–
Liriodendron tulipifiera	–	–	–	–	2	–	–	1	–	<0.5	–	–
Rhododendron calendulaceum	–	–	–	–	–	1	–	1	4	–	–	–
Craya glabra	–	–	–	–	–	4	<0.5	2	6	5	–	–
Carya tomentosa	–	–	–	–	–	–	–	2	–	–	–	–
Carya ovalis	–	–	–	–	–	–	–	<0.5	–	–	–	–
Nyssa sylvatica	–	–	1	–	–	–	2	4	1	2	7	–
Oxydendrum arboreum	–	–	<0.5	1	–	1	3	8	14	16	1	1
Castanea dentata	–	–	–	–	2	5	7	9	10	12	1	–
Sassafras albidum	–	–	–	–	–	1	1	1	1	4	<0.5	–
Quercus alba	–	–	–	–	–	2	1	8	24	10	<0.5	–
Robinia pseudoacacia	–	–	–	–	–	4	5	1	3	8	3	<0.5
Quercus prinus	–	–	–	–	–	3	4	15	4	16	11	1
Quercus veluntina	–	–	–	–	–	–	<0.5	<0.5	1	1	–	–
Quercus coccinea	–	–	–	–	–	–	1	–	–	–	–	1
Pinus rigida	–	–	–	–	–	–	–	7	1	1	11	46
Pinus pungens	–	–	–	–	–	–	–	–	1	4	54	49

Table 9
Frequencies of tree presence percentage by station number, in ordinal scale.

Ordinal scale	0	1	2	3	4
Tree presence	No data recorded	≤0.5%	≤1%	≤8%	>8%
Frequency (x_{ij})	285	30	68	65	44

based on an equitable frequency percentage for each category. Table 9 summarizes the frequencies of tree presence data for this new ordinal scale with 5 categories. Apart from the first category, which is for sites and tree species without presences recorded, the categories with the highest frequencies are 2 and 3 (tree presence percentages between 0.5% and 8%).

Here it is important to remark that we defined another ordinal scale with six categories in the beginning of the data analysis. The current category 3 was split in two subcategories (from 1% to 2% and from 2% to 8%) in that former ordinal scale. However, models fitted to these data indicated that the corresponding estimated score parameters ϕ_k for those two adjacent categories were very close to each other. If $\phi_k = \phi_{k+1}$ then the adjacent category logit between those two categories, say k and $k+1$ is

$$\log\left(\frac{P\left[y_{ij}=k+1 \mid i\in r\right]}{P\left[y_{ij}=k \mid i\in r\right]}\right) = (\mu_{k+1}-\mu_k) + (\phi_{k+1}-\phi_k)(\alpha_r+\beta_j+\gamma_{rj})$$
$$= \mu_{k+1}-\mu_k.$$

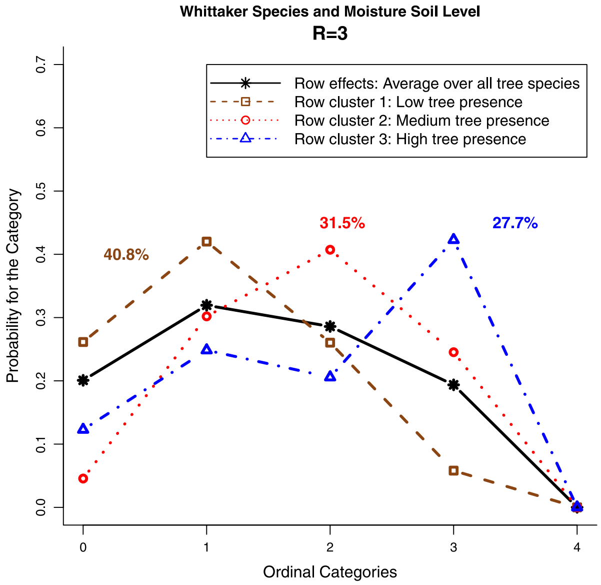


Fig. 9. $R = 3$ tree presence groups. The lines depict the probability for the category $\widehat{\theta}_{r_i jk} = P\left[y_{ij} = k \mid i \in r\right]$ (see Eq. (5)) for each group r and the average over all students (black line). The percentage labeling is the estimated posteriori probability $\widehat{\pi}_r$ that member of each group r will respond to questions in each ordinal category (Eq. (18)).

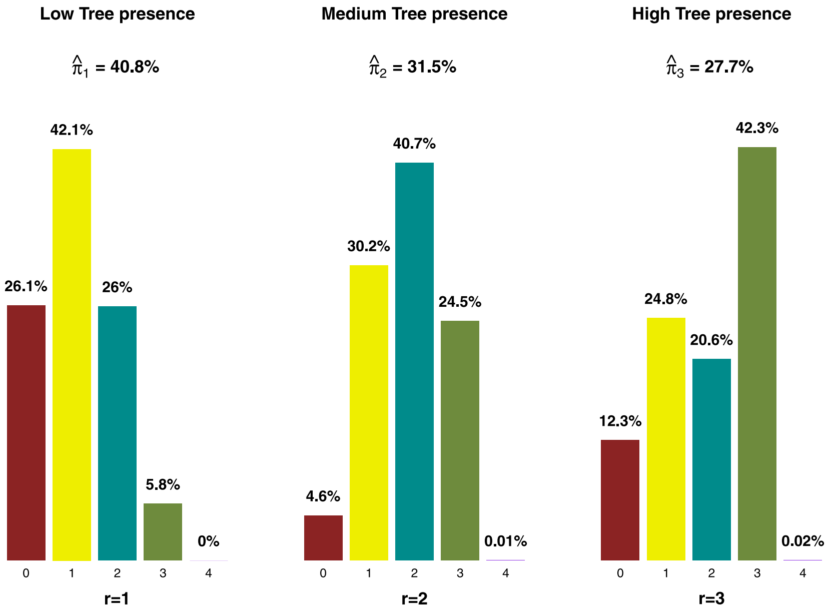


Fig. 10. $R = 3$ tree presence profiles. The percentages represent the probability $\widehat{\theta}_{rk}$ in each category (Eq. (5)).

This implies that the relative frequencies in these two categories are independent of the clustering structure. Therefore retaining the distinction between k and $k+1$ is not informative about the clustering structure. In that case, the model still holds with the same scores if the ordinal scale is collapsed by combining those two adjacent categories into one single response category. Since we regard the $\{\mu_k\}$ as nuisance parameters, this collapsing does not limit our inference in any way. However, they should not be collapsed if keeping the original ordinal categories entails an ecological interest. Therefore, the dimensions of the data matrix Y with the responses are $n = 41$ tree species and $m = 12$ site stations where each observation can take one of the 5 possible categories described above.

After fitting a complete set of models and comparing them by using information criteria (see the summarized results in Table G.19 in Appendix G), the selected model (either using AIC or AIC$_c$) was the stereotype model version including row clustering with $R = 3$ row groups and with interaction factors $(\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj}))$. Figs. 9 and 10 show the profiles for the three resultant row clusters. For instance, Fig. 9 depicts that the highest probability for the row cluster number 3 (showed with a line with diamond symbols) is in ordinal category 3 and Fig. 10 shows a first set of bars where the highest probabilities are in the categories with tree presence below 1%. Therefore, tree species classified in the first row group are those with a lower level of presence.

6. Discussion

We have introduced a set of likelihood models based on the ordinal stereotype model and have introduced fuzzy clustering via finite mixtures in order to reduce the dimensionality of the problem and simplify the interpretation. Furthermore, we have described a procedure to derive the maximum likelihood estimators of the parameters of the suite of developed models by using the EM algorithm. Computation can be slow in this stage, though we have substantially reduced the time required by calling **C** from **R**. In addition, we have tested model comparison setting up a simulation study conducted a study in which the results show that AIC, AIC$_c$ and AIC$_u$ are consistent information criteria to score fitted models based on our likelihood-based finite mixture model approach for ordinal data sets. In particular, AIC performs best among the tested information criteria to select the model with the correct number of clusters in a wide range of scenarios. Finally, we have demonstrated our approach by means of two examples with real data and have tested the reliability of our methodology through a simulation study. We have detected that there is an indication of multimodality of the likelihood surface in the column clustering model. A strategy to deal with this is to implement a convergence strategy where several starting values are tested over the parameter vector in order to obtain the global maximum.

There are numerous applications for these likelihood-based clustering models in fields where multivariate techniques are necessary. The advantage of our approach is its likelihood-based foundation because maximum likelihood theory provides estimators and model selection.

Many count or percentage data sets have extreme variabilities, for example very high counts and very low counts in ecological community data. Replacing these high counts with "medium" and "high" ordinal categories makes the actual counts less influential in the model fitting, giving broad categories which enable us to detect major overall patterns.

This approach can be enriched adding covariates if they are available. As the mixture models have an unknown number of components, a Bayesian approach implementing a reversible jump MCMC sampler is also feasible (cf. Arnold et al., 2010). Another possible future interesting development might be to assume random effects for indicated row and column effects (α_i or β_j) to reduce the number of unknown parameters.

We have 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 $\{\phi_{jk}\}$.

An important future development would be new data visualization methods for finite mixtures models based on the stereotype model. In Figs. 7–10 we have displayed the different group profiles using $\{\overline{\phi}_{(i.)}\}$ from the data. New visualization tools may lead to the generation of new hypotheses in data exploration and the identification of patterns in the data. Output from these models allows us to determine a new spacing of the ordinal categories, dictated by the data. This will lead to more informative visualization, for example with the equal spacing of categories along the x-axis in Figs. 9 and 10 replaced with the more appropriate spacing determined by the fitted parameter values. The estimation of the spacing among ordinal responses in our methodology is an improvement over other ordinal data models such as proportional odds model and continuation-ratio model although more research in performance comparison with others equivalent methods is needed. Finally, another research direction to explore would be to compare clustering structures resulting from our methodology with those with binary, count or continuous data. In particular for data sets which original data is not ordinal and we apply transformations like in the Example 2 (Section 5.2.2).

Acknowledgments

The authors are sincerely grateful to Dr. Ivy Liu and Dr. Eleni Matechou for making many valuable suggestions and stimulating discussions about this work. We are grateful to two referees and an associate editor for helpful comments and references. Daniel Fernández acknowledges Victoria University of Wellington for providing financial support through Victoria Doctoral Scholarship program.

Appendix A. Response probabilities in the ordered stereotype model

In this appendix, we describe the relationship between models (3) and (4), which were formulated in Section 2.2.

From Eq. (4) with the linear predictor including the covariates and $\sum_{k=1}^{q} P\left[y_{ij} = k \mid \boldsymbol{x}\right] = 1$, we calculate

$$P\left[y_{ij} = 1 \mid \boldsymbol{x}\right]\left(1 + \sum_{\ell=2}^{q} \exp(\mu_\ell + \phi_\ell \boldsymbol{\delta}'\boldsymbol{x})\right) = 1.$$

As $\mu_1 = \phi_1 = 0$ for identifiability reasons, then

$$P\left[y_{ij} = 1 \mid \boldsymbol{x}\right] = \frac{1}{\sum_{\ell=1}^{q} \exp(\mu_\ell + \phi_\ell \boldsymbol{\delta}'\boldsymbol{x})}. \tag{A.1}$$

Therefore, Eq. (3) can be obtained from (4) just using the above expression (A.1) for $P\left[y_{ij} = 1 \mid \boldsymbol{x}\right]$.

Table D.10
Parameter configuration for 5 tested scenarios in the row clustering case.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
	$m = 5$	$m = 5$	$m = 5$	$m = 5$	$m = 10$
$R = 2$	$\pi_1 = 0.450$	$\pi_1 = 0.450$	$\pi_1 = 0.950$	$\pi_1 = 0.450$	$\pi_1 = 0.450$
	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$
	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$
	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$
	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.500$	$\phi_2 = 0.335$
	$\phi_3 = 0.662$	$\phi_3 = 0.972$	$\phi_3 = 0.662$	$\phi_3 = 0.500$	$\phi_3 = 0.662$
	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$
	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$
	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$
	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$
	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$
$R = 3$	$\pi_1 = 0.200$	$\pi_1 = 0.200$	$\pi_1 = 0.470$	$\pi_1 = 0.200$	$\pi_1 = 0.200$
	$\pi_2 = 0.500$	$\pi_2 = 0.500$	$\pi_2 = 0.050$	$\pi_2 = 0.500$	$\pi_2 = 0.500$
	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$
	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$
	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$
	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.500$	$\phi_2 = 0.335$
	$\phi_3 = 0.662$	$\phi_3 = 0.972$	$\phi_3 = 0.662$	$\phi_3 = 0.500$	$\phi_3 = 0.662$
	$\alpha_1 = 3.634$	$\alpha_1 = 3.634$	$\alpha_1 = 3.634$	$\alpha_1 = 3.634$	$\alpha_1 = 3.634$
	$\alpha_2 = -0.819$	$\alpha_2 = -0.819$	$\alpha_2 = -0.819$	$\alpha_2 = -0.819$	$\alpha_2 = -0.819$
	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$
	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$
	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$
	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$
$R = 4$	$\pi_1 = 0.150$	$\pi_1 = 0.150$	$\pi_1 = 0.310$	$\pi_1 = 0.150$	$\pi_1 = 0.150$
	$\pi_2 = 0.300$	$\pi_2 = 0.300$	$\pi_2 = 0.050$	$\pi_2 = 0.300$	$\pi_2 = 0.300$
	$\pi_3 = 0.250$	$\pi_3 = 0.250$	$\pi_3 = 0.320$	$\pi_3 = 0.250$	$\pi_3 = 0.250$
	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$	$\mu_2 = 0.814$
	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$	$\mu_3 = 0.951$
	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$	$\mu_4 = 0.207$
	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.500$	$\phi_2 = 0.335$
	$\phi_3 = 0.662$	$\phi_3 = 0.972$	$\phi_3 = 0.662$	$\phi_3 = 0.500$	$\phi_3 = 0.662$
	$\alpha_1 = 3.634$	$\alpha_1 = 3.634$	$\alpha_1 = 3.634$	$\alpha_1 = 3.634$	$\alpha_1 = 3.634$
	$\alpha_2 = -0.819$	$\alpha_2 = -0.819$	$\alpha_2 = -0.819$	$\alpha_2 = -0.819$	$\alpha_2 = -0.819$
	$\alpha_3 = 2.911$	$\alpha_3 = 2.911$	$\alpha_3 = 2.911$	$\alpha_3 = 2.911$	$\alpha_3 = 2.911$
	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$	$\beta_1 = -0.427$
	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$	$\beta_2 = 1.285$
	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$	$\beta_3 = 1.872$
	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$	$\beta_4 = -0.097$

Notes: $\mu_1 = 0, \phi_1 = 0, \phi_4 = 1$ for all the scenarios.
$\beta_5 = 2.20, \beta_6 = 3.00, \beta_7 = -2.00, \beta_8 = 3.90$ and $\beta_9 = -3.50$ in Scenario 5.

Appendix B. EM algorithm formulae. Column clustering

In Section 3, we described the model fitting procedure for the row clustering case. In this appendix, the fitting procedure is formulated for the case of column clustering.

The latent variable relating to the missing information for the actual membership of the columns is X_{jc}. The posterior probabilities of membership once we have observed the data $\{y_{ij}\}$ are $\widehat{X}_{jc}$ and the set of *a priori* probabilities are $\{\kappa_c\}$. Ω is the parameter vector for the case of column clustering. For the M-step, we use the sum-to-zero constraints on each row and column of the γ iteration matrix and on the column effect parameters $\{\beta_c\}$ ($\sum_c \beta_c = 0$) in order to avoid identifiability problems.

The column clustering model-specific formulae of EM-algorithm follow.

E-step:

$$\widehat{X}_{jc}^{(t)} = \frac{\widehat{\kappa}_c^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} \left(\widehat{\theta}_{ick}^{(t-1)}\right)^{I(y_{ij}=k)}}{\sum_{\ell=1}^{C} \left\{\widehat{\kappa}_\ell^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} \left(\widehat{\theta}_{i\ell k}^{(t-1)}\right)^{I(y_{ij}=k)}\right\}}$$

and

$$\widehat{Q}(\Omega \mid \Omega^{(t-1)}) = \sum_{j=1}^{m} \sum_{c=1}^{C} \widehat{X}_{jc}^{(t)} \log(\widehat{\kappa}_c^{(t-1)}) + \sum_{i=1}^{n} \sum_{i=1}^{m} \sum_{k=1}^{q} \sum_{c=1}^{C} \widehat{X}_{jc}^{(t)} I(y_{ij} = k) \log\left(\widehat{\theta}_{ick}^{(t-1)}\right).$$

Table D.11
Parameter configuration for 5 tested scenarios in the biclustering case.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
	$m = 5$	$m = 5$	$m = 5$	$m = 5$	$m = 10$
$R = 2, C = 2$	$\pi_1 = 0.450$	$\pi_1 = 0.450$	$\pi_1 = 0.450$	$\pi_1 = 0.450$	$\pi_1 = 0.450$
	$\kappa_1 = 0.450$	$\kappa_1 = 0.450$	$\kappa_1 = 0.950$	$\kappa_1 = 0.450$	$\kappa_1 = 0.450$
	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$
	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$
	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$
	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.500$	$\phi_2 = 0.335$
	$\phi_3 = 0.672$	$\phi_3 = 0.972$	$\phi_3 = 0.672$	$\phi_3 = 0.500$	$\phi_3 = 0.672$
	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$
	$\beta_1 = 0.777$	$\beta_1 = 0.777$	$\beta_1 = 0.777$	$\beta_1 = 0.777$	$\beta_1 = 0.777$
$R = 2, C = 3$	$\pi_1 = 0.450$	$\pi_1 = 0.450$	$\pi_1 = 0.450$	$\pi_1 = 0.450$	$\pi_1 = 0.450$
	$\kappa_1 = 0.200$	$\kappa_1 = 0.200$	$\kappa_1 = 0.470$	$\kappa_1 = 0.200$	$\kappa_1 = 0.200$
	$\kappa_2 = 0.500$	$\kappa_2 = 0.500$	$\kappa_2 = 0.050$	$\kappa_2 = 0.500$	$\kappa_2 = 0.500$
	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$
	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$
	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$
	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.500$	$\phi_2 = 0.335$
	$\phi_3 = 0.672$	$\phi_3 = 0.972$	$\phi_3 = 0.672$	$\phi_3 = 0.500$	$\phi_3 = 0.672$
	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$
	$\beta_1 = -2.128$	$\beta_1 = -2.128$	$\beta_1 = -2.128$	$\beta_1 = -2.128$	$\beta_1 = -2.128$
	$\beta_2 = 3.212$	$\beta_2 = 3.212$	$\beta_2 = 3.212$	$\beta_2 = 3.212$	$\beta_2 = 3.212$
$R = 3, C = 2$	$\pi_1 = 0.200$	$\pi_1 = 0.200$	$\pi_1 = 0.200$	$\pi_1 = 0.200$	$\pi_1 = 0.200$
	$\pi_2 = 0.500$	$\pi_2 = 0.500$	$\pi_2 = 0.500$	$\pi_2 = 0.500$	$\pi_2 = 0.500$
	$\kappa_1 = 0.450$	$\kappa_1 = 0.450$	$\kappa_1 = 0.950$	$\kappa_1 = 0.450$	$\kappa_1 = 0.450$
	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$
	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$
	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$
	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.500$	$\phi_2 = 0.335$
	$\phi_3 = 0.672$	$\phi_3 = 0.972$	$\phi_3 = 0.672$	$\phi_3 = 0.500$	$\phi_3 = 0.672$
	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$
	$\alpha_2 = 3.251$	$\alpha_2 = 3.251$	$\alpha_2 = 3.251$	$\alpha_2 = 3.251$	$\alpha_2 = 3.251$
	$\beta_1 = 0.777$	$\beta_1 = 0.777$	$\beta_1 = 0.777$	$\beta_1 = 0.777$	$\beta_1 = 0.777$
$R = 3, C = 3$	$\pi_1 = 0.200$	$\pi_1 = 0.200$	$\pi_1 = 0.200$	$\pi_1 = 0.200$	$\pi_1 = 0.200$
	$\pi_2 = 0.500$	$\pi_2 = 0.500$	$\pi_2 = 0.500$	$\pi_2 = 0.500$	$\pi_2 = 0.500$
	$\kappa_1 = 0.200$	$\kappa_1 = 0.200$	$\kappa_1 = 0.47$	$\kappa_1 = 0.20$	$\kappa_1 = 0.200$
	$\kappa_2 = 0.500$	$\kappa_2 = 0.500$	$\kappa_2 = 0.050$	$\kappa_2 = 0.500$	$\kappa_2 = 0.500$
	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$	$\mu_2 = 0.914$
	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$	$\mu_3 = 0.511$
	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$	$\mu_4 = 0.107$
	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.335$	$\phi_2 = 0.500$	$\phi_2 = 0.335$
	$\phi_3 = 0.672$	$\phi_3 = 0.972$	$\phi_3 = 0.672$	$\phi_3 = 0.500$	$\phi_3 = 0.672$
	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$	$\alpha_1 = 1.634$
	$\alpha_2 = 3.251$	$\alpha_2 = 3.251$	$\alpha_2 = 3.251$	$\alpha_2 = 3.251$	$\alpha_2 = 3.251$
	$\beta_1 = -2.128$	$\beta_1 = -2.128$	$\beta_1 = -2.128$	$\beta_1 = -2.128$	$\beta_1 = -2.128$
	$\beta_2 = 3.212$	$\beta_2 = 3.212$	$\beta_2 = 3.212$	$\beta_2 = 3.212$	$\beta_2 = 3.212$

Notes: $\mu_1 = 0, \phi_1 = 0, \phi_4 = 1$ for all the scenarios.

M-step:

$$\widehat{\kappa}_c^{(t)} = \frac{1}{m} \sum_{j=1}^{m} \left(\frac{\widehat{\kappa}_c^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} \left(\widehat{\theta}_{ick}^{(t-1)} \right)^{I(y_{ij}=k)}}{\sum_{l=1}^{C} \left\{ \widehat{\kappa}_l^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} \left(\widehat{\theta}_{ilk}^{(t-1)} \right)^{I(y_{ij}=k)} \right\}} \right)$$

and

$$\max_{\Omega} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{c=1}^{C} \widehat{X}_{jc} I(y_{ij} = k) \log \left(\widehat{\theta}_{ick} \right) \right],$$

conditional on the identifiability constraints on the parameters.

Appendix C. EM algorithm formulae. Biclustering

In Section 3, we described the model fitting procedure for the row clustering case. In this appendix, the fitting procedure is formulated for the case of biclustering.

Table E.12
Simulation study (Section 5.1). Estimated score parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j} + \boldsymbol{\gamma_{rj}})$. MLEs and their standard errors from the score and row membership parameters ($\{\phi_k\}$, $\{\pi_r\}$) for different number of row clusters R and sample sizes n are shown.

R	Numpar	True param.	$n = 200$		$n = 500$		$n = 1000$		$n = 5000$	
			Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
2	15	$\phi_2 = 0.335$	0.324	0.235	0.336	0.140	0.337	0.098	0.336	0.044
		$\phi_3 = 0.672$	0.655	0.206	0.674	0.123	0.672	0.087	0.671	0.038
		$\pi_1 = 0.550$	0.556	0.063	0.542	0.035	0.550	0.024	0.554	0.010
3	21	$\phi_2 = 0.335$	0.372	0.236	0.321	0.142	0.331	0.100	0.339	0.069
		$\phi_3 = 0.672$	0.709	0.165	0.668	0.102	0.678	0.074	0.675	0.052
		$\pi_1 = 0.200$	0.201	0.091	0.219	0.015	0.172	0.008	0.202	0.007
		$\pi_2 = 0.500$	0.353	0.148	0.487	0.114	0.451	0.031	0.491	0.015
4	27	$\phi_2 = 0.335$	0.373	0.236	0.374	0.144	0.348	0.093	0.345	0.049
		$\phi_3 = 0.672$	0.727	0.167	0.771	0.095	0.692	0.070	0.682	0.031
		$\pi_1 = 0.200$	0.084	0.129	0.099	0.118	0.179	0.101	0.181	0.059
		$\pi_2 = 0.350$	0.327	0.201	0.401	0.141	0.334	0.128	0.346	0.022
		$\pi_3 = 0.230$	0.196	0.181	0.259	0.128	0.179	0.099	0.214	0.059
5	33	$\phi_2 = 0.335$	0.323	0.243	0.374	0.154	0.335	0.102	0.335	0.073
		$\phi_3 = 0.672$	0.698	0.152	0.744	0.097	0.684	0.071	0.675	0.050
		$\pi_1 = 0.200$	0.151	0.128	0.214	0.107	0.212	0.051	0.209	0.003
		$\pi_2 = 0.120$	0.114	0.155	0.136	0.121	0.128	0.061	0.121	0.003
		$\pi_3 = 0.230$	0.210	0.186	0.224	0.130	0.228	0.057	0.234	0.008
		$\pi_4 = 0.300$	0.462	0.198	0.440	0.157	0.388	0.111	0.311	0.011
6	39	$\phi_2 = 0.335$	0.442	0.238	0.404	0.147	0.333	0.103	0.346	0.071
		$\phi_3 = 0.672$	0.741	0.166	0.766	0.106	0.709	0.075	0.680	0.056
		$\pi_1 = 0.150$	0.181	0.172	0.167	0.121	0.131	0.081	0.138	0.012
		$\pi_2 = 0.300$	0.182	0.155	0.221	0.091	0.225	0.058	0.227	0.009
		$\pi_3 = 0.100$	0.091	0.161	0.081	0.102	0.087	0.077	0.093	0.014
		$\pi_4 = 0.200$	0.246	0.139	0.166	0.081	0.194	0.044	0.194	0.005
		$\pi_5 = 0.150$	0.235	0.166	0.191	0.118	0.178	0.099	0.165	0.012

The latent variables relating to the missing information for the actual membership of the rows and columns are Z_{ir} and X_{jc} respectively. The posterior probabilities of membership once we have observed the data $\{y_{ij}\}$ are $\widehat{Z}_{ir}$ for the rows and $\widehat{X}_{jc}$ for the columns. The set of *a priori* probabilities are $\{\pi_r\}$ (rows) and $\{\kappa_c\}$ (columns). Ω is the parameter vector for the case of biclustering. For the M-step, we use the sum-to-zero constraints on each row and column of the γ iteration matrix and on row effect parameters $\{\alpha_r\}$ and column effect parameters $\{\beta_c\}$ ($\sum_r \alpha_r = \sum_c \beta_c = 0$) in order to avoid identifiability problems. The biclustering model-specific formulae of EM-algorithm follow (see the detailed formulation of the biclustering model by Pledger and Arnold, 2014).

E-step:

$$\widehat{Z}_{ir}^{(t)} = \frac{\widehat{\pi}_r^{(t-1)} \prod_{j=1}^{m} \prod_{k=1}^{q} \left\{ \sum_{c=1}^{C} \widehat{\kappa}_c \left(\widehat{\theta}_{rck}^{(t-1)} \right)^{I(y_{ij}=k)} \right\}}{\sum_{\ell=1}^{R} \widehat{\pi}_\ell^{(t-1)} \prod_{j=1}^{m} \prod_{k=1}^{q} \left\{ \sum_{c=1}^{C} \widehat{\kappa}_c \left(\widehat{\theta}_{\ell ck}^{(t-1)} \right)^{I(y_{ij}=k)} \right\}}$$

and

$$\widehat{X}_{jc}^{(t)} = \frac{\widehat{\kappa}_c^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} \left\{ \sum_{r=1}^{R} \widehat{\pi}_r \left(\widehat{\theta}_{rck}^{(t-1)} \right)^{I(y_{ij}=k)} \right\}}{\sum_{\ell=1}^{C} \widehat{\kappa}_\ell^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} \left\{ \sum_{r=1}^{R} \widehat{\pi}_r \left(\widehat{\theta}_{r\ell k}^{(t-1)} \right)^{I(y_{ij}=k)} \right\}}.$$

The E-step of the EM algorithm calls for the expected value of the complete data log-likelihood taking into account the fact that the only data unknown is $\{z_{ir}\}$ and $\{x_{jc}\}$ conditional on the observed data $\{y_{ij}\}$:

$$\widehat{Q}(\Omega \mid \Omega^{(t-1)}) = \sum_{i=1}^{n} \sum_{r=1}^{R} \log\left(\widehat{\pi}_r^{(t-1)}\right) E\left[z_{ir} \mid \{y_{ij}\}, \Omega^{(t-1)}\right] + \sum_{j=1}^{m} \sum_{c=1}^{C} \log\left(\widehat{\kappa}_c^{(t-1)}\right) E\left[x_{jc} \mid \{y_{ij}\}, \Omega^{(t-1)}\right]$$
$$+ \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{r=1}^{R} \sum_{c=1}^{C} I(y_{ij} = k) \log\left(\widehat{\theta}_{rck}^{(t-1)}\right) E\left[z_{ir} x_{jc} \mid \{y_{ij}\}, \Omega^{(t-1)}\right].$$

Table E.13
Simulation study (Section 5.1). Estimated score parameters for stereotype model including biclustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\alpha_r + \boldsymbol{\beta_c} + \boldsymbol{\gamma_{rc}})$. MLEs and their standard errors from the score, row and column membership parameters ($\{\phi_k\}, \{\pi_r\}, \{\kappa_c\}$) for different number of row and column clusters R and C and sample sizes n are shown.

R	C	Numpar	True param.	$n = 25$		$n = 50$		$n = 100$	
				Mean	S.E.	Mean	S.E.	Mean	S.E.
2	2	10	$\phi_2 = 0.335$	0.383	0.304	0.346	0.178	0.339	0.138
			$\phi_3 = 0.672$	0.705	0.260	0.699	0.232	0.678	0.118
			$\pi_1 = 0.600$	0.604	0.173	0.583	0.107	0.601	0.078
			$\kappa_1 = 0.400$	0.366	0.178	0.407	0.097	0.402	0.076
2	3	13	$\phi_2 = 0.335$	0.391	0.325	0.342	0.183	0.337	0.140
			$\phi_3 = 0.672$	0.696	0.294	0.659	0.197	0.669	0.099
			$\pi_1 = 0.600$	0.628	0.188	0.591	0.087	0.604	0.061
			$\kappa_1 = 0.400$	0.412	0.171	0.398	0.102	0.400	0.088
			$\kappa_2 = 0.200$	0.189	0.168	0.201	0.094	0.199	0.059
3	2	13	$\phi_2 = 0.335$	0.298	0.299	0.341	0.131	0.336	0.111
			$\phi_3 = 0.672$	0.713	0.297	0.693	0.166	0.675	0.109
			$\pi_1 = 0.300$	0.288	0.176	0.304	0.101	0.303	0.077
			$\pi_2 = 0.400$	0.371	0.163	0.388	0.099	0.397	0.065
			$\kappa_1 = 0.400$	0.421	0.181	0.401	0.137	0.400	0.111
3	3	17	$\phi_2 = 0.335$	0.401	0.313	0.388	0.201	0.347	0.131
			$\phi_3 = 0.672$	0.701	0.277	0.669	0.181	0.671	0.093
			$\pi_1 = 0.300$	0.325	0.182	0.312	0.106	0.304	0.066
			$\pi_2 = 0.400$	0.371	0.178	0.381	0.101	0.397	0.071
			$\kappa_1 = 0.400$	0.384	0.157	0.398	0.092	0.402	0.063
			$\kappa_2 = 0.200$	0.219	0.148	0.210	0.104	0.195	0.061

Table E.14
Simulation study (Section 5.1). Estimated score parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j} + \boldsymbol{\gamma_{rj}})$ when $\phi_2 = \phi_3$ or π_2 is small. MLEs and their standard errors from the score and row membership parameters ($\{\phi_k\}, \{\pi_r\}$) for different number of row clusters R and sample sizes n are shown.

R	Numpar	True param.	$n = 200$		$n = 500$		$n = 1000$		$n = 5000$	
			Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
2	15	$\boldsymbol{\phi_2 = 0.500}$	0.492	0.221	0.483	0.131	0.501	0.091	0.498	0.041
		$\boldsymbol{\phi_3 = 0.500}$	0.524	0.089	0.502	0.056	0.511	0.036	0.503	0.017
		$\pi_1 = 0.550$	0.595	0.060	0.572	0.036	0.525	0.027	0.551	0.011
3	21	$\boldsymbol{\phi_2 = 0.500}$	0.495	0.217	0.484	0.133	0.492	0.093	0.498	0.040
		$\boldsymbol{\phi_3 = 0.500}$	0.525	0.456	0.504	0.256	0.509	0.146	0.511	0.094
		$\pi_1 = 0.200$	0.202	0.097	0.180	0.013	0.171	0.010	0.203	0.009
		$\pi_2 = 0.500$	0.495	0.140	0.512	0.087	0.504	0.040	0.453	0.012
4	27	$\boldsymbol{\phi_2 = 0.500}$	0.498	0.212	0.520	0.083	0.517	0.067	0.506	0.055
		$\boldsymbol{\phi_3 = 0.500}$	0.545	0.242	0.491	0.116	0.524	0.063	0.513	0.025
		$\pi_1 = 0.200$	0.196	0.165	0.188	0.102	0.193	0.058	0.197	0.016
		$\pi_2 = 0.350$	0.416	0.181	0.406	0.125	0.373	0.042	0.375	0.012
		$\pi_3 = 0.230$	0.240	0.262	0.285	0.163	0.249	0.021	0.280	0.009
3	21	$\phi_2 = 0.335$	0.332	0.228	0.336	0.096	0.334	0.068	0.341	0.047
		$\phi_3 = 0.672$	0.666	0.207	0.674	0.088	0.661	0.064	0.682	0.045
		$\pi_1 = 0.400$	0.344	0.052	0.419	0.031	0.414	0.018	0.422	0.012
		$\boldsymbol{\pi_2 = 0.015}$	0.010	0.123	0.024	0.065	0.012	0.042	0.019	0.026

The expectations in the former two terms are simply $\widehat{Z}_{ir}$ and $\widehat{X}_{jc}$. However, the lack of *a posteriori* independence of the $\{z_{ir}\}$ and $\{x_{jc}\}$ makes the evaluation of $E\left[z_{ir}x_{jc} \mid \{y_{ij}\}, \Omega\right]$ computationally expensive as it requires a sum either over all possible allocations of rows to row groups, or over all possible allocations of columns to column groups.

The variational approximation employed by Govaert and Nadif (2005) is a solution to this problem:

$$E\left[z_{ir}x_{jc} \mid \{y_{ij}\}, \Omega\right] \simeq E\left[z_{ir} \mid \{y_{ij}\}, \Omega\right] E\left[x_{jc} \mid \{y_{ij}\}, \Omega\right] = \widehat{Z}_{ir}\widehat{X}_{jc}.$$

In that manner, the E-step of the EM algorithm is approximated as:

$$\begin{aligned}\widehat{Q}(\Omega \mid \Omega^{(t-1)}) = {} & \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{Z}_{ir}\log\left(\widehat{\pi}_r^{(t-1)}\right) + \sum_{j=1}^{m}\sum_{c=1}^{C}\widehat{X}_{jc}\log\left(\widehat{\kappa}_c^{(t-1)}\right)\\ & + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{r=1}^{R}\sum_{c=1}^{C}\widehat{Z}_{ir}\widehat{X}_{jc}I(y_{ij}=k)\log\left(\widehat{\theta}_{rck}^{(t-1)}\right).\end{aligned} \tag{C.1}$$

Table E.15
Simulation study (Section 5.1). Estimated score parameters for stereotype model including column clustering $\boldsymbol{\mu_k + \phi_k(\alpha_i + \beta_c + \gamma_{ic})}$ when $\phi_2 = \phi_3$ or κ_2 is small and $m = 15$. MLEs and their standard errors from the score and column membership parameters ($\{\phi_k\}, \{\kappa_c\}$) for different number of column clusters C, number of columns m and sample sizes n are shown.

C	True param.	$n = 25$			$n = 50$		
		Numpar	Mean	S.E.	Numpar	Mean	S.E.
2	$\boldsymbol{\phi_2} = \mathbf{0.700}$	31	0.776	0.100	56	0.706	0.056
	$\boldsymbol{\phi_3} = \mathbf{0.700}$		0.882	0.111		0.856	0.076
	$\kappa_1 = 0.400$		0.382	0.121		0.409	0.097
3	$\boldsymbol{\phi_2} = \mathbf{0.700}$	33	0.761	0.123	58	0.734	0.068
	$\boldsymbol{\phi_3} = \mathbf{0.700}$		0.796	0.111		0.768	0.056
	$\kappa_1 = 0.400$		0.411	0.105		0.423	0.045
	$\kappa_2 = 0.200$		0.176	0.077		0.200	0.035
3	$\phi_2 = 0.335$	33	0.328	0.209	58	0.362	0.080
	$\phi_3 = 0.672$		0.713	0.157		0.633	0.140
	$\kappa_1 = 0.400$		0.401	0.170		0.373	0.086
	$\boldsymbol{\kappa_2} = \mathbf{0.015}$		0.026	0.149		0.027	0.084

Table E.16
Simulation study (Section 5.1). Estimated score parameters for stereotype model including biclustering $\mu_k + \phi_k(\alpha_r + \beta_c + \gamma_{rc})$ when $\phi_2 = \phi_3$ or π_2 and κ_2 are small. MLEs and their standard errors from the score, row and column membership parameters ($\{\phi_k\}, \{\pi_r\}, \{\kappa_c\}$) for different number of row and column clusters R and C and sample sizes n are shown.

R	C	Numpar	True param.	$n = 25$		$n = 50$		$n = 100$	
				Mean	S.E.	Mean	S.E.	Mean	S.E.
2	2	10	$\boldsymbol{\phi_2} = \mathbf{0.500}$	0.477	0.361	0.507	0.164	0.522	0.141
			$\boldsymbol{\phi_3} = \mathbf{0.500}$	0.602	0.388	0.593	0.208	0.529	0.174
			$\pi_1 = 0.600$	0.573	0.273	0.540	0.110	0.638	0.103
			$\kappa_1 = 0.400$	0.367	0.289	0.406	0.168	0.401	0.062
2	3	13	$\boldsymbol{\phi_2} = \mathbf{0.500}$	0.659	0.346	0.620	0.141	0.539	0.078
			$\boldsymbol{\phi_3} = \mathbf{0.500}$	0.664	0.281	0.612	0.205	0.629	0.088
			$\pi_1 = 0.600$	0.514	0.284	0.689	0.149	0.670	0.146
			$\kappa_1 = 0.400$	0.363	0.401	0.410	0.188	0.358	0.089
			$\kappa_2 = 0.200$	0.253	0.311	0.249	0.142	0.237	0.029
3	2	13	$\boldsymbol{\phi_2} = \mathbf{0.500}$	0.628	0.247	0.429	0.164	0.544	0.065
			$\boldsymbol{\phi_3} = \mathbf{0.500}$	0.651	0.230	0.576	0.137	0.546	0.060
			$\pi_1 = 0.300$	0.297	0.225	0.278	0.136	0.268	0.037
			$\pi_2 = 0.400$	0.408	0.228	0.409	0.130	0.362	0.059
			$\kappa_1 = 0.400$	0.482	0.285	0.418	0.143	0.409	0.088
3	3	17	$\boldsymbol{\phi_2} = \mathbf{0.500}$	0.404	0.210	0.497	0.106	0.447	0.036
			$\boldsymbol{\phi_3} = \mathbf{0.500}$	0.563	0.212	0.558	0.110	0.518	0.039
			$\pi_1 = 0.300$	0.340	0.127	0.317	0.045	0.307	0.014
			$\pi_2 = 0.400$	0.358	0.149	0.396	0.105	0.385	0.014
			$\kappa_1 = 0.400$	0.458	0.141	0.382	0.108	0.399	0.016
			$\kappa_2 = 0.200$	0.239	0.085	0.219	0.076	0.204	0.013
2	3	13	$\phi_2 = 0.335$	0.390	0.320	0.338	0.141	0.301	0.057
			$\phi_3 = 0.672$	0.760	0.271	0.620	0.107	0.642	0.080
			$\pi_1 = 0.400$	0.382	0.142	0.488	0.100	0.423	0.090
			$\kappa_1 = 0.400$	0.457	0.136	0.479	0.185	0.402	0.079
			$\boldsymbol{\kappa_2} = \mathbf{0.015}$	0.013	0.089	0.014	0.064	0.018	0.018
3	2	13	$\phi_2 = 0.335$	0.326	0.223	0.356	0.156	0.332	0.076
			$\phi_3 = 0.672$	0.691	0.307	0.618	0.146	0.613	0.079
			$\pi_1 = 0.400$	0.463	0.180	0.373	0.068	0.397	0.024
			$\boldsymbol{\pi_2} = \mathbf{0.015}$	0.028	0.194	0.019	0.078	0.020	0.055
			$\kappa_1 = 0.400$	0.403	0.113	0.385	0.070	0.408	0.033
3	3	17	$\phi_2 = 0.335$	0.386	0.256	0.320	0.125	0.331	0.063
			$\phi_3 = 0.672$	0.685	0.221	0.631	0.140	0.674	0.080
			$\pi_1 = 0.400$	0.391	0.170	0.311	0.098	0.415	0.068
			$\boldsymbol{\pi_2} = \mathbf{0.015}$	0.025	0.159	0.021	0.079	0.017	0.043
			$\kappa_1 = 0.400$	0.445	0.188	0.386	0.079	0.398	0.038
			$\boldsymbol{\kappa_2} = \mathbf{0.015}$	0.019	0.130	0.014	0.043	0.022	0.015

Table F.17
List of 10 questions of applied statistics course feedback forms. Each question was written so that "agree" indicates a positive view of the course.

Questions		
Q1. The way this course was organised has helped me to learn.		
Disagree	Neither agree nor disagree	Agree
□	□	□
Q2. Important course information-such as learning objectives, deadlines, assessments and grading criteria-was communicated clearly.		
Disagree	Neither Agree nor Disagree	Agree
□	□	□
Q3. Preparing for the assessments has helped me to learn.		
Disagree	Neither Agree nor Disagree	Agree
□	□	□
Q4. Comments and feedback I received during the course have helped me learn more effectively.		
Disagree	Neither Agree nor Disagree	Agree
□	□	□
Q5. This course encouraged me to think critically.		
Disagree	Neither Agree nor Disagree	Agree
□	□	□
Q6. This course encouraged me to think creatively.		
Disagree	Neither Agree nor Disagree	Agree
□	□	□
Q7. This course has helped me to develop my communication skills.		
Disagree	Neither Agree nor Disagree	Agree
□	□	□
Q8. This course has stimulated my interest in learning more about this subject.		
Disagree	Neither Agree nor Disagree	Agree
□	□	□
Q9. I value highly what I have learned from this course.		
Disagree	Neither Agree nor Disagree	Agree
□	□	□
Q10. Overall, I would rate the quality of this course as very good:		
Disagree	Neither Agree nor Disagree	Agree
□	□	□

M-step:

$$\widehat{\kappa}_c^{(t)} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\widehat{\kappa}_c^{(t-1)}\prod_{i=1}^{n}\prod_{k=1}^{q}\left(\widehat{\theta}_{ick}^{(t-1)}\right)^{I(y_{ij}=k)}}{\sum_{l=1}^{C}\left\{\widehat{\kappa}_l^{(t-1)}\prod_{i=1}^{n}\prod_{k=1}^{q}\left(\widehat{\theta}_{ilk}^{(t-1)}\right)^{I(y_{ij}=k)}\right\}}\right)$$

and

$$\widehat{\pi}_r^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{\pi}_r^{(t-1)}\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\widehat{\theta}_{rjk}^{(t-1)}\right)^{I(y_{ij}=k)}}{\sum_{l=1}^{R}\left\{\widehat{\pi}_l^{(t-1)}\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\widehat{\theta}_{ljk}^{(t-1)}\right)^{I(y_{ij}=k)}\right\}}\right)$$

and

$$\max_{\Omega}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{r=1}^{R}\sum_{c=1}^{C} I(y_{ij}=k)\log\left(\widehat{\theta}_{rck}\right)\widehat{Z}_{ir}^{(t)}\widehat{X}_{jc}^{(t)}\right],$$

conditional on the identifiability constraints on the parameters and assume independence between $\widehat{Z}_{ir}$ and $\widehat{X}_{jc}$.

The variational approximation presents several drawbacks (see e.g. Keribin et al., 2012 for a discussion on this topic). In our work, we have not employed the variational approximation for the ultimate MLEs. Instead, we have used an alternative

Table F.18
Applied Statistics feedback form responses. 70 students (rows), 10 questions (Q1–Q10) and 3 categories for each question: "disagree" (coded as 1), "neither agree or disagree" (coded as 2) and "agree" (coded as 3).

Students ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	1	1	1	1	2	2	2	2	1	1
2	1	1	1	1	3	3	3	3	1	1
3	1	1	1	2	1	2	3	2	2	2
4	2	2	1	1	2	2	3	3	2	2
5	1	1	1	1	2	3	3	3	1	3
6	3	2	1	2	1	3	3	3	2	3
7	1	1	1	3	2	2	3	1	1	1
8	1	1	1	1	1	2	2	1	2	1
9	2	1	1	1	2	3	2	3	1	2
10	1	1	1	1	2	2	2	2	2	2
11	1	1	1	1	1	1	1	1	1	1
12	2	1	1	3	3	3	3	2	1	3
13	2	1	1	2	3	3	3	1	1	2
14	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	2	1	1
16	1	1	1	2	2	1	3	2	1	2
17	1	1	1	2	2	2	2	2	1	2
18	1	1	1	1	1	2	2	1	1	2
19	1	1	1	1	2	2	1	1	1	1
20	2	2	1	2	1	2	2	2	1	2
21	1	1	1	1	1	2	2	1	1	1
22	3	3	1	1	1	1	2	1	1	1
23	1	1	1	1	1	2	2	1	1	1
24	1	1	1	1	1	2	2	2	1	1
25	1	1	1	1	1	2	2	2	1	1
26	1	1	2	2	1	1	2	3	1	2
27	1	1	1	1	1	1	2	1	1	1
28	1	1	1	1	1	1	1	1	1	1
29	1	1	1	2	2	2	2	2	2	2
30	1	1	1	2	2	3	3	2	1	2
31	1	1	1	1	1	1	1	2	1	1
32	1	1	1	1	1	2	1	1	1	1
33	2	1	1	2	1	2	2	2	1	2
34	1	1	1	2	2	2	2	2	2	2
35	1	2	1	1	2	2	2	3	1	2
36	1	1	1	2	1	2	1	2	2	2
37	1	1	1	2	2	2	2	2	1	2
38	1	1	1	2	2	2	2	1	1	1
39	1	1	1	1	1	1	1	2	1	2
40	1	1	2	2	1	3	3	1	2	2
41	1	1	1	1	1	3	1	3	1	1
42	1	1	1	1	2	2	2	1	1	1
43	1	1	1	2	2	3	3	3	2	2
44	2	1	1	2	2	3	2	3	1	1
45	3	1	1	2	3	3	3	1	1	3
46	1	1	1	1	2	2	2	2	1	2
47	1	1	1	1	1	2	2	3	1	1
48	1	1	1	2	1	1	2	1	1	2
49	1	1	1	2	1	2	1	2	1	1
50	1	1	1	1	2	2	2	3	2	2
51	1	1	1	1	2	3	3	2	2	2
52	1	1	1	1	1	2	2	3	2	1
53	1	1	1	1	1	1	1	1	1	1
54	2	3	1	3	3	3	3	3	3	2
55	1	1	1	2	1	1	1	2	1	1
56	1	1	1	1	1	1	1	1	1	1
57	1	1	1	2	2	3	2	2	1	1
58	1	1	1	1	1	2	2	2	1	1
59	1	1	1	1	1	1	2	1	2	1
60	2	1	1	2	3	3	3	3	2	2
61	1	1	1	1	1	1	2	1	1	2
62	1	1	1	1	2	2	2	1	1	1
63	1	1	1	2	1	2	2	3	1	2
64	1	1	1	1	3	3	3	3	1	1
65	1	1	1	1	1	1	3	1	1	1
66	1	1	1	2	1	2	2	2	1	2
67	3	1	1	1	1	1	1	3	1	3

(*continued on next page*)

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004

Table F.18 (*continued*)

Students ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
68	1	1	1	1	2	2	2	3	1	1
69	1	3	1	1	2	2	2	2	1	1
70	3	1	1	3	3	3	1	3	1	2

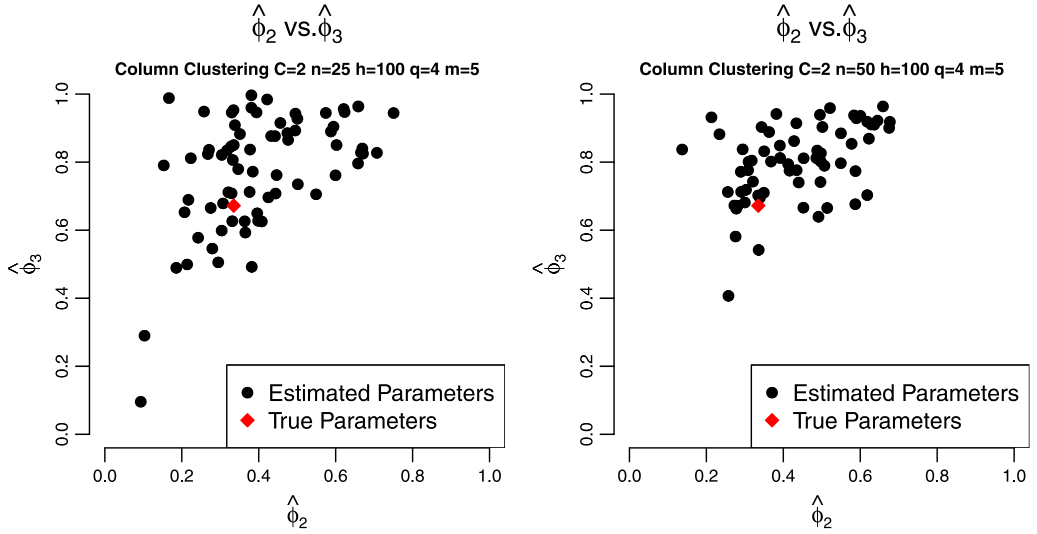


Fig. E.11. Simulation study (Section 5.1): convergence of $\widehat{\phi}_2$ and $\widehat{\phi}_3$ for the stereotype model including the column clustering $(\alpha_i + \beta_c)$ with $C = 2$ column clusters. n, h, q, m describe the sample size, the number of replicates, the number of categories and the number of covariates respectively. The diamond point represents the true value of the parameter.

procedure with the aim of ensuring a solution avoiding approximation. Thus, the MLEs from the EM algorithm are used as starting points in order to numerically maximize the incomplete-data log-likelihood (12) (or (13)).

Appendix D. Model comparison. Simulations study

Tables D.10 and D.11 summarize the parameter configuration for each scenario in the row clustering and biclustering cases.

Appendix E. Simulations. Other scenarios

Tables E.12 and E.13 summarize the results for the simulation scenarios (Section 5.1) including the interaction factors for row clustering and biclustering version respectively. Figs. E.11 and E.12 show the evolution of the convergence for the estimated score parameters $\widehat{\phi}_2$ and $\widehat{\phi}_3$ in the case with $R = 2$ row groups with $C = 2$ column groups and biclustering with $R = 2$ and $C = 2$ row and column groups, respectively. Tables E.14–E.16 show two particular scenarios described in Section 5.1 for the row clustering, column clustering and biclustering respectively.

Appendix F. Applied statistics course feedback

The list of questions are shown in Table F.17 and the data set with the responses of 70 students over 10 questions giving feedback about a second year Applied Statistics course is given in Table F.18.

Appendix G. Tree presences in Great Smoky Mountains

Table G.19 summarize the suite of fitted models for R.H. Whittaker's study of vegetation of the Great Smoky Mountains data set (Whittaker, 1956, Table 3). For each model, the information criteria AIC, AIC$_c$, BIC and ICL-BIC were computed.

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004

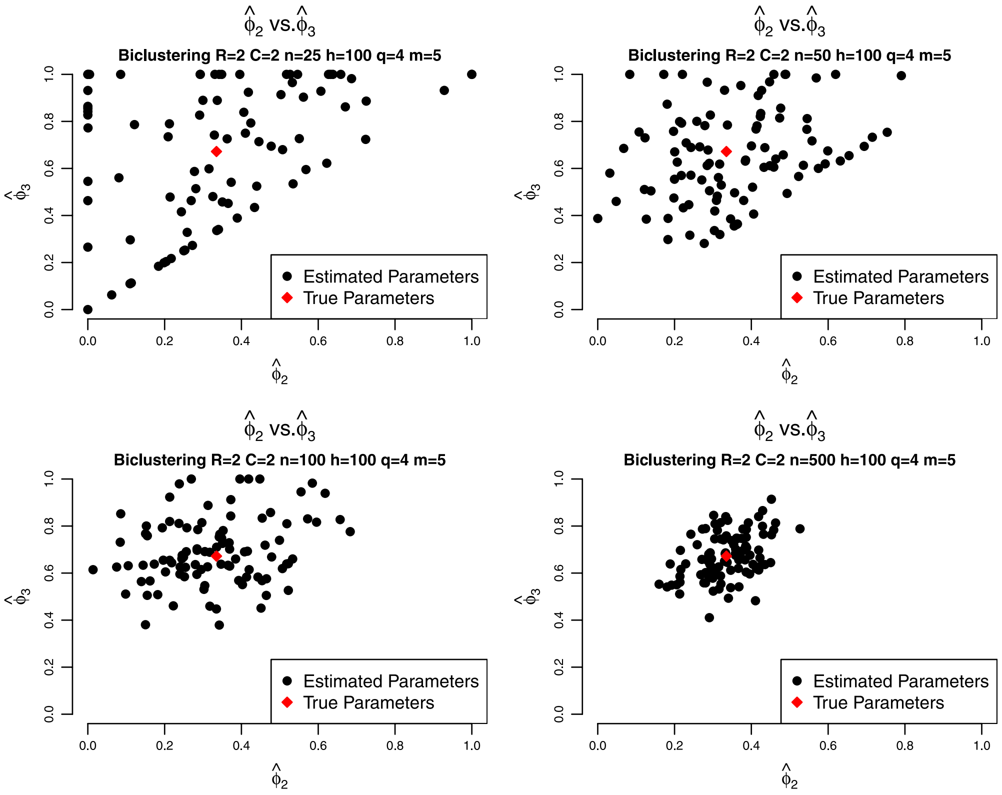


Fig. E.12. Simulation study (Section 5.1): convergence of $\widehat{\phi_2}$ and $\widehat{\phi_3}$ for the stereotype model including the biclustering $(\alpha_r + \beta_c)$ with $R = 2$ row and $C = 2$ column clusters. n, h, q, m describe the sample size, the number of replicates, the number of categories and the number of covariates respectively. The diamond point represents the true value of the parameter.

Table G.19
Suite of models fitted for R.H. Whittaker's study of vegetation. For each information criterion, the best model in each group (no clustering, row clustering, column clustering and biclustering) is shown in boldface.

Model		R	C	npar	AIC	AIC$_c$	BIC	ICL-BIC
Null model	μ_k	1	1	7	671.41	671.70	700.79	700.79
Row effects	$\mu_k + \phi_k\alpha_i$	n	1	47	572.15	582.77	769.48	769.48
Column effects	$\mu_k + \phi_k\beta_j$	1	m	18	581.09	582.70	**656.67**	**656.67**
Main effects	$\mu_k + \phi_k(\alpha_i + \beta_j)$	n	m	58	**544.22**	**560.61**	787.83	787.83
Row clustering	$\mu_k + \phi_k\alpha_r$	2	1	9	549.10	549.56	**586.88**	**605.18**
		3	1	11	553.05	553.70	599.24	617.32
		4	1	13	556.94	557.82	611.52	630.75
	$\mu_k + \phi_k(\alpha_r + \beta_j)$	2	m	20	555.65	557.62	639.62	655.14
		3	m	22	558.91	561.27	651.28	671.11
		4	m	24	563.56	566.35	664.32	712.84
	$\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj})$	2	m	31	534.06	538.66	664.21	669.38
		3	$\boldsymbol{m}$	**44**	**518.01**	**527.30**	702.75	712.35
		4	m	57	529.27	545.07	768.58	779.42
Column clustering	$\mu_k + \phi_k\beta_c$	1	2	9	**549.10**	**549.56**	**586.88**	**605.18**
		1	3	11	570.25	570.90	616.43	699.76
	$\mu_k + \phi_k(\alpha_i + \beta_c)$	n	2	49	580.88	592.45	646.61	703.41
		n	3	51	594.38	606.93	648.50	679.20
Biclustering	$\mu_k + \phi_k(\alpha_r + \beta_c)$	2	2	11	**551.49**	**552.14**	**597.67**	**621.13**
		3	2	13	580.14	581.02	634.72	698.37
		2	3	13	581.17	582.12	634.89	697.80
		3	3	15	584.14	585.29	647.12	712.45
	$\mu_k + \phi_k(\alpha_r + \beta_c + \gamma_{rc})$	2	2	12	553.49	554.25	603.87	625.57
		3	2	15	586.65	587.80	619.63	655.07
		2	3	15	559.49	560.63	622.46	657.83
		3	3	19	569.77	571.55	649.54	676.11

References

Agresti, A., 2002. Categorical Data Analysis, second ed. In: Wiley Series in Probability and Statistics, Wiley-Interscience.

Agresti, A., 2010. Analysis of Ordinal Categorical Data, second ed. In: Wiley Series in Probability and Statistics, Wiley.

Agresti, A., Lang, J.B., 1993. Quasi-symmetric latent class models, with application to rater agreement. Biometrics 49, 131–139.

Ahn, J., Mukherjee, B., Banerjee, M., Cooney, K.A., 2009. Bayesian inference for the stereotype regression model: application to a case-control study of prostate cancer. Stat. Med. 28, 3139–3157.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), 2nd International Symposium on Information Theory, pp. 267–281.

Anderson, J.A., 1984. Regression and ordered categorical variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 46, 1–30.

Arnold, R., Hayakawa, Y., Yip, P., 2010. Capture–recapture estimation using finite mixtures of arbitrary dimension. Biometrics 66, 644–655.

Banfield, J.D., Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803–821.

Biernacki, C., Celeux, G., Govaert, G., 1998. Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. Technical Report 3521. INRIA, Rhône-Alpes.

Biernacki, C., Celeux, G., Govaert, G., 1999. An improvement of the nec criterion for assessing the number of clusters in mixture model. Pattern Recognit. Lett. 267–272.

Biernacki, C., Govaert, G., 1997. Using the classification likelihood to choose the number of clusters. Comput. Sci. Stat. 451–457.

Bock, R., Jones, L., 1968. The Measurement and Prediction of Judgment and Choice. In: Holden-Day Series in Psychology, Holden-Day.

Böhning, D., Seidel, W., Alfò, M., Garel, B., Patilea, V., Walther, G., 2007. Advances in mixture models. Comput. Statist. Data Anal. 51, 5205–5210.

Bozdogan, H., 1987. Model selection and Akaikes's information criterion (AIC): The general theory and its analytical extensions. Psycometrika 52, 345–370.

Bozdogan, H., 1994. Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, vol. 2: Multivariate Statistical Modeling. Kluwer Academic Publishers, the Netherlands, Dordrecht, pp. 69–113.

Breen, R., Luijkx, R., 2010. Assessing proportionality in the proportional odds model for ordinal logistic regression. Sociol. Methods Res. 39, 3–24.

Burnham, K., Anderson, D., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, second ed. Springer.

Chen, L.C., Yu, P.S., Tseng, V.S., 2011. A weighted fuzzy-based biclustering method for gene expression data. Int. J. Data Min. Bioinform. 5, 89–109.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39, 1–38.

DeSantis, S.M., Houseman, E.A., Coull, B.A., Stemmer-Rachamimov, A., Betensky, R.A., 2008. A penalized latent class model for ordinal data. Biostatistics 9, 249–262.

DeSarbo, W.S., Fong, D.K.H., Liechty, J., Kim Saxton, M., 2004. A hierarchical Bayesian procedure for two-mode cluster analysis. Psychometrika 69, 547–572.

Everitt, B.S., Leese, M., Landau, S., 2001. Cluster Analysis, fourth ed. Hodder Arnold Publication.

Figueredo, M.A.T., Jain, A.K., 2002. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 24, 1–16.

Fonseca, J.R.S., Cardoso, M., 2007. Mixture-model cluster analysis using information theoretical criteria. Intell. Data Anal. 11, 155–173.

Goodman, L.A., 1979. Simple models for the analysis of association in cross-classifications having ordered categories. J. Amer. Statist. Assoc. 74, 537–552.

Gotelli, N.J., Graves, G.R., 1996. Null Models in Ecology. Smithsonian Institution Press, Washington D.C.

Govaert, G., Nadif, M., 2005. An EM algorithm for the block mixture model. IEEE Trans. Pattern Anal. Mach. Intell. 27, 643–647.

Govaert, G., Nadif, M., 2010. Latent block model for contingency table. Comm. Statist. Theory Methods 39, 416–425.

Greenland, S., 1994. Alternative models for ordinal logistic regression. Stat. Med. 13, 1665–1677.

Hennig, C., Liao, T.F., 2013. How to find an appropriate clustering for mixed type variables with application to socioeconomic stratification. J. Roy. Statist. Sci. Ser. C Appl. Statist. 62, 309–369.

Hoffman, D.L., Franke, G.R., 1986. Correspondence analysis: graphical representation of categorical data in marketing research. J. Mark. Res. 23, 213–227.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Jobson, J.D., 1992. Applied Multivariate Data Analysis: Categorical and Multivariate Methods. In: Springer Texts in Statistics, Springer.

Johnson, S.C., 1967. Hierarchical clustering schemes. Psychometrika 2, 241–254.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data an Introduction to Cluster Analysis. Wiley, New York.

Keribin, C., Brault, V., Celeux, G., Govart, G., 2012. Estimation and Selection for the Latent Block Model on Categorical Data. Technical Report. INRIA Research Report.

Kuss, O., 2006. On the estimation of the stereotype regression model. Comput. Statist. Data Anal. 50, 1877–1890.

Labiod, L., Nadif, M., 2011. Co-clustering for binary and categorical data with maximum modularity. In: ICDM, pp. 1140–1145.

Lewis, S.J.G., Foltynie, T., Blackwell, A.D., Robbins, T.W., Owen, A.M., Barker, R.A., 2003. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. J. Neurol. Neurosurg. Psychiatry. 76, 343–348.

Liu, I., Agresti, A., 2005. The analysis of ordered categorical data: an overview and a survey of recent developments. TEST: Official J. Spanish Soc. Statist. Oper. Res. 14, 1–73.

Manly, B.F.J., 2005. Multivariate Statistical Methods: A Primer. Chapman & Hall/CRC Press, Boca Raton, FL.

Manly, B.F.J., 2007. Randomization, Bootstrap and Monte Carlo Methods in Biology, third ed. Chapman & Hall, London.

Matechou, E., Liu, I., Pledger, S., Arnold, R., 2011. Biclustering models for ordinal data. In: Presentation at the NZ Statistical Assn. Annual Conference, University of Auckland, 28–31 August 2011.

McCullagh, P., 1980. Regression models for ordinal data. J. Roy. Statist. Soc. 42, 109–142.

McCune, B., Grace, J.B., 2002. Analysis of Ecological Communities. Volume 28. MjM Software Design.

McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl. Stat. 318–324.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. In: Statistics, Textbooks and Monographs, M. Dekker.

McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. In: Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. In: Wiley Series in Probability and Statistics.

McPartland, D., Gormley, I., 2013. Clustering ordinal data via latent variable models. In: Algorithms from and for Nature and Life. In: Studies in Classification, Data Analysis, and Knowledge Organization, Springer International Publishing.

McQuarrie, A., Shumway, R., Tsai, C.L., 1997. The model selection criterion AICu. Statist. Probab. Lett. 34, 285–292.

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

Moustaki, I., 2000. A latent variable model for ordinal variables. Appl. Psychol. Meas. 211–233.

Mukherjee, B., Ahn, J., Liu, I., Rathouz, P.J., Sanchez, B., 2008. Fitting stratified proportional odds models by amalgamating conditional likelihoods. Stat. Med. 27, 4950–4971.

Pledger, S., 2000. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. Biometrics 56, 434–442.

Pledger, S., Arnold, R., 2014. Multivariate methods using mixtures: correspondence analysis, scaling and pattern-detection. Comput. Statist. Data Anal. URL: http://dx.doi.org/10.1016/j.csda.2013.05.013.

Quinn, G.P., Keough, M.J., 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press.

R Development Core Team 2010. R: A Language and Environment for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org.ISBN3-900051-07-0.

Rocci, R., Vichi, M., 2008. Two-mode multi-partitioning. Comput. Statist. Data Anal. 52, 1984–2003.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Snell, E.J., 1964. A scaling procedure for ordered categorical data. Biometrics 20, 592–607.

Stahl, D., Sallis, H., 2012. Model-based cluster analysis. In: Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 4. pp. 341–358.

Stevens, S., 1946. On the theory of scales of measurement. Science 103, 677–680.

Vermunt, J.K., 2001. The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. Appl. Psychol. Meas. 25, 283–294.

Vichi, M., 2001. Double *k*-means clustering for simultaneous classification of objects and variables. In: Borra, S., Rocci, R., Vichi, M., Schader, M. (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization. Springer, pp. 43–52.

Whittaker, R.H., 1956. Vegetation of the great smoky mountains. Ecol. Monograph 26, 1–80.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14, 1–37.

Wu, H.M., Tzeng, S., Chen, C.H., 2007. Matrix visualization. In: Handbook of Data Visualization. pp. 681–708.

Please cite this article in press as: Fernández, D., et al., Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.11.004