***Published paper*:** Fernandez, D. [et al.]. Finite mixture biclustering of discrete type multivariate data. "Advances in data analysis and classification", 15 Maig 2018, vol. 13, núm. 1, p. 117-143. doi:10.1007/s11634-018-0324-3

Adv Data Anal Classif
https://doi.org/10.1007/s11634-018-0324-3

REGULAR ARTICLE

Finite mixture biclustering of discrete type multivariate data

Daniel Fernandez[1,2] · **Richard Arnold**[2] · **Shirley Pledger**[2] · **Ivy Liu**[2] · **Roy Costilla**[3]

Received: 29 November 2016 / Revised: 21 February 2018 / Accepted: 2 May 2018

Abstract Many of the methods which deal with clustering in matrices of data are based on mathematical techniques such as distance-based algorithms or matrix decomposition and eigenvalues. In general, it is not possible to use statistical inferences or select the appropriateness of a model via information criteria with these techniques because there is no underlying probability model. This article summarizes some recent model-based methodologies for matrices of binary, count, and ordinal data, which are modelled under a unified statistical framework using finite mixtures to group the rows and/or columns. The model parameter can be constructed from a linear predictor of parameters and covariates through link functions. This likelihood-based one-mode and two-mode fuzzy clustering provides maximum likelihood estimation of parameters and the options of using likelihood information criteria for model comparison. Additionally, a Bayesian approach is presented in which the parameters and the number of clusters are estimated simultaneously from their joint posterior distribution. Visualization tools focused on ordinal data, the fuzziness of the clustering structures, and analogies of various standard plots used in the multivariate analysis are presented. Finally, a set of future extensions is enumerated. 1

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11634-018-0324-3) contains supplementary material, which is available to authorized users.

✉ Daniel Fernandez
df.martinez@pssjd.org

1 Institut de Recerca Sant Joan de Déeu, Parc Sanitari Sant Joan de Déeu, CIBERSAM, Dr. Antoni Pujades, 42, 08830 Sant Boi de Llobregat, Barcelona, Spain

2 School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

3 Institute for Molecular Bioscience, University of Queensland, Saint Lucia, Australia

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

Keywords Classification · EM algorithm · Fuzzy clustering · Mixture models · Ordinal data · RJMCMC · Visualisation tools

Mathematics Subject Classification 62F15 · 62F86 · 62H12 · 62H30 · 62H86

1 Introduction

Cluster analysis has been widely used in many areas such as ecology, marketing, and computer science to identify groups, patterns, or clusters in a data set. For example, we may have n individuals completing a health questionnaire containing m questions, with y_{ij} being the response of person i to question j. We thus have data in an $n \times m$ array Y, along with other possible covariates. We may wish to find groups of persons (rows) each containing individuals with similar patterns of responses, and simultaneously find groups of correlated questions (columns). This leads to a two-mode clustering, or a biclustering problem.

In general, there are non-model-based and model-based approaches for cluster analysis. The most common heuristic non-model-based approach uses a criterion (Friedman and Rubin 1967) on the sum of within-cluster sums of squares, e.g., k-means clustering (MacQueen 1967; Hartigan and Wong 1979; Jobson 1992; Vichi 2001; McCune and Grace 2002; Rocci and Vichi 2008), where the data points are iteratively moved from one cluster to another until there is no improvement in the criterion. In addition, many metric methods have been developed including hierarchical clustering, multidimensional scaling, association analysis, correspondence analysis and ordination [see e.g. Johnson (1967), Manly (2005), Everitt et al. (2011), Quinn and Keough (2002)]. Although these methods have been successful in solving many practical problems, no statistical inference is available because they are not based on statistical likelihoods. Statistical tests can only be constructed through the use of resampling methods (Manly 2007; Gotelli and Graves 1996), but it is still not clear how to decide the number of clusters (Fraley and Raftery 1998).

A long-standing model-based approach to clustering assumes the data come from a mixture of probability distributions [see e.g., McLachlan and Basford (1988), McLachlan and Peel (2000), Everitt et al. (2011), Böhning et al. (2007), Wu et al. (2008), Melnykov and Maitra (2010), Melnykov (2013), Matechou et al. (2016)]. For continuous outcomes y_{ij} the clustering methodology is based on multivariate normal mixtures and the estimation is usually carried out using the expectation-maximization (EM) algorithm (Dempster et al. 1977). This approach provides a probability clustering where each subject is probabilistically classified across the groups, allowing a richer description of the data than a method that definitively allocates each observation to a single cluster. In this setting we might classify one individual definitively into Group 1, another definitively into Group 2, but a third might have 80%/20% membership probabilities for these groups.

The model-based approach has some distinct advantages over the non-model based approaches listed above. In particular, it allows the use of statistical inference and information criteria (Akaike 1973; Hurvich and Tsai 1989; Schwarz 1978; Biernacki et al. 1998) to compare models in order to select a suitable number of clusters. Additionally, it allows an accurate representation and inference of complex distributions, identification of different groups, better handling of missing data, and the possibility to fit structured data (e.g. longitudinal data) (McLachlan and Peel 2000). On the other hand, model based clustering is computationally intensive when implemented using the EM algorithm or Bayesian methods. Moreover, finding a good starting point for the EM algorithm is not easy, which is a common issue for finite mixture models. With a bad choice, the parameter estimates might reach a local maximum of the likelihood. Additionally, and unlike metric methods (e.g., k-means clustering), practitioners need to have some basic knowledge on statistical models. Most metric methods are more user friendly to solve many practical problems.

It is only in recent times that the model-based clustering of non-continuous outcomes has received significant attention, and the clustering of such data is the subject of this paper. Specifically, we discuss the likelihood-based biclustering of arrays of non-continuous data, where each of n individuals has a set of m binary, count, or ordinal measurements. These types of data are common across many disciplines. Examples include incidence and abundance matrices in ecological communities where the rows are species and the columns are samples, and binary/ordinal item response analysis with respondents in the rows and questions in the columns. The cluster analysis of ordinal data has received remarkably little attention in the literature, and such data are often treated as continuous in order to apply existing methodologies.

This paper reviews our recent work in this area. Pledger (2000) and Arnold et al. (2010) proposed biclustering using mixtures for binary data. Pledger and Arnold (2014) developed an approach via finite mixtures for binary and count data using basic Bernoulli or Poisson building blocks. This approach unified a suite of models, some new and some previously published proposals for binary data and count data (Govaert and Nadif 2003, 2010; Nadif and Govaert 2005), and showed that new geometric insights provide likelihood-based analogues of multidimensional scaling, association analysis, correspondence analysis, pattern detection, ordination and biplots. Hui et al. (2015) compared single-mode clustering via finite mixtures with using normally-distributed random effects, for Poisson and negative binomial models. For ordinal data, Matechou et al. (2016), Fernández et al. (2016), Fernández and Pledger (2016), and Fernández and Arnold (2016) developed and applied clustering models for ordinal data using the assumption of proportional odds (McCullagh 1980) or the ordered stereotype model (Anderson 1984). Our work bears some similarity to latent class models (Goodman 1974; Haberman 1979; McCutcheon 1987) in the sense that the models consist of sets of subjects with unobserved homogeneous response distributions (Agresti and Lang 1993; Moustaki 2000; Vermunt 2001; DeSantis et al. 2008; Breen and Luijkx 2010; McParland and Gormley 2013). Nevertheless, our models have the flexibility across row, column and biclustering for the data in an $n \times m$ array with or without covariates. In our work fuzzy allocation of rows and columns to corresponding clusters is usually achieved by performing the EM algorithm or by Bayesian methods. In addition, the fuzzy clustering approach allows novel data visualization tools for depicting the results of the clustering.

This paper is structured as follows. Section 2 contains definitions of the models and their formulation using fuzzy clustering via finite mixtures. Model fitting by using the iterative EM algorithm and a Bayesian approach are described in Sect. 3. Graphical

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

displays for ordinal and count data are presented in Sect. 4, and we conclude with a discussion, technical notes, and extensions in Sect. 5. A "Supplementary Appendix" contains a summary of the definitions of all information criteria measures used in the paper (Sect. S1), an outline of the Reversible-jump MCMC algorithm and of the relabelling procedure to overcome the label switching problem (Sects. S2, S3, respectively), how average scores for graphical displaying of ordinal data are computed (Sect. S4), details on the data set used throughout this paper and on a new graphical tool for ordinal data based on mosaic plots (Sects. S5, S6), and technical details (Sect. S7).

2 Finite mixture models

The widespread use of finite mixture models as a mathematical-based method for statistical modeling of unknown random phenomena in an extremely flexible way has increased over the last 20 years (McLachlan and Peel 2000). An appropriate choice of the components that make up the finite mixture model allows both the accurate representation of complex distributions and inference about the random phenomena observed.

Finite mixture modeling can be viewed as latent variable analysis with a latent categorical variable describing the group or subpopulation membership, and the latent classes being described by the different components of the mixture distribution (Skrondal and Rabe-Hesketh 2004).

In the setting of an $n \times m$ matrix of observations $Y = \{y_{ij}\}$ we may wish to cluster the rows, the columns, or both simultaneously (biclustering). Here we give expressions for row clustering and biclustering. Results for column clustering follow straightforwardly by exchanging rows and columns in the row clustered case.

The data we use throughout this paper is the *student feedback form* ordinal data set (Fernández et al. 2016). It has the responses of 70 students giving feedback about an applied statistics course. 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 list of questions and the data set are given in Tables S4 and S5 in "Supplementary Appendix S5".

2.1 The row-clustered model

In row clustering we assume that each m-dimensional row y_i $(i = 1, \ldots, n)$ is a realization drawn from the R component finite mixture

$$f(y_i|x_i, \Omega) = \sum_{r=1}^{R} \pi_r f_r(y_i|x_i, \theta_r).$$

Here x_i is a $d \times 1$ set of covariates, $(\pi_1, \ldots, \pi_R)$ are the mixture component probabilities, and θ_r is the set of parameters corresponding to the rth mixture component

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

$f_r(y_i|x_i, \theta_r)$. Ω contains all the unknown parameters in the mixture, $\{(\pi_r, \theta_r)\}_{r=1}^R$. The mixing probabilities π_r satisfy

$$\sum_{r=1}^{R} \pi_r = 1, \quad 0 \leq \pi_r \leq 1, \quad r = 1, \ldots, R,$$

and π_r is the a priori probability that a row in the matrix belongs to mixture component r. We write $i \in r$ to indicate the event that row i is drawn from mixture component r.

The individual mixture component distributions $f_r(y_i|x_i, \theta_r)$ are the probability densities/mass functions of y_i given $i \in r$. These distributions may be specified distinctly, or may be members of a single family of distributions—differing only through their dependence on x_i and θ_r. If so then the subscript r on $f_r(\cdot|\cdot)$ is redundant, and we have $f_r(y_i|x_i, \theta_r) = f(y_i|x_i, \theta_r)$.

A further simplification occurs when the m elements of y_i are conditionally independent given x_i and θ_r, so that

$$f(y_i|x_i, \theta_r) = \prod_{j=1}^{m} f(y_{ij}|x_{ij}, \theta_{rj}) \qquad \text{if} \quad i \in r$$

with x_{ij} a $d_j \times 1$ subset of x_i. Most of the models we discuss are of this form, however there are important extensions for repeated measures and other correlated data settings which we discuss briefly in Sect. 5.

The likelihood of the full $n \times m$ data array sums over all possible allocations of the n rows to the R clusters:

$$L(\Omega|\{y_{ij}, x_{ij}\}) = \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \cdots \pi_{r_n} \prod_{i=1}^{n} \prod_{j=1}^{m} f(y_{ij}|x_{ij}, \theta_{r_i j}),$$

which can be simplified to

$$L(\Omega|\{y_{ij}, x_{ij}\}) = \prod_{i=1}^{n} \left[\sum_{r=1}^{R} \pi_r \prod_{j=1}^{m} f(y_{ij}|x_{ij}, \theta_{r_i j}) \right]. \tag{1}$$

In the case of the student feedback form data set, row clustering implies the clustering of students and not questions. Additionally, the model formulation for column clustering is similar, with clustering of columns but not rows, i.e. clustering of questions but not students.

Maximisation of expressions such as (1) is analytically complex and numerically demanding, and the EM algorithm is often used to find parameter estimates. In the mixture setting it is convenient to introduce the $R \times 1$ latent group membership variable Z_i with $Z_{ir} = 1$ if $i \in r$ and $Z_{ir'} = 0$ for $r' \neq r$. A priori the group memberships follow a multinomial distribution

$$Z_i = (Z_{i1}, \ldots, Z_{iR})^T \sim \text{Multinomial}(1; \pi_1, \ldots, \pi_R)$$

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

with $\sum_{r=1}^{R} Z_{ir} = 1$. These group memberships form the missing data when estimation is carried out using the EM Algorithm (see Sect. 3 below). The joint distribution of (y_i, Z_i) is then

$$f(y_i, Z_i|x_i, \{\theta_r\}) = \prod_{r=1}^{R} [\pi_r f(y_i|x_i, \theta_r)]^{Z_{ir}}$$

leading to the complete data likelihood

$$L_c(\Omega|\{y_{ij}\}, \{Z_{ir}\}) = \prod_{i=1}^{n}\prod_{j=1}^{m}\prod_{r=1}^{R} [\pi_r f(y_i|x_i, \theta_r)]^{Z_{ir}}$$

which is much more amenable to maximisation due to its product structure.

The a posteriori distribution of Z_i is multinomial

$$Z_i = (Z_{i1}, \ldots, Z_{iR})^T|Y \sim \text{Multinomial}(1; \widehat{Z}_1, \ldots, \widehat{Z}_R).$$

Here $\widehat{Z}_{ir} = P[i \in r|Y]$ is the estimated probability, conditional on the data, that observation i comes from group r.

2.2 The biclustered model

Simultaneous clustering of both rows and columns, also known as biclustering, allocates each row to one of R row groups, and each column to one of C column groups. The notation of the row clustered model is augmented as follows. The a priori probability that column j is in group c (written $j \in c$) is κ_c so that the mixture distribution, assuming full conditional independence of every cell from every other, is

$$f(y_{ij}|x_{ij}, \Omega) = \sum_{r=1}^{R} \pi_r \sum_{c=1}^{C} \kappa_c f(y_{ij}|x_{ij}, \theta_{rc}) \quad \text{for} \quad i = 1, \ldots, n, \quad j = 1, \ldots, m$$

with x_{ij} a $d_j \times 1$ subset of x_i.

The likelihood sums over all possible allocations of rows to R clusters and columns to C clusters:

$$L(\Omega|\{y_{ij}, x_{ij}\}) = \sum_{c_1=1}^{C} \cdots \sum_{c_m=1}^{C} \kappa_{c_1} \cdots \kappa_{c_m} \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \cdots \pi_{r_n} \prod_{i=1}^{n}\prod_{j=1}^{m} f(y_{ij}|x_{ij}, \theta_{r_i c_j})$$

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

which can be simplified to

$$L(\Omega|\{y_{ij}\}) = \sum_{c_1=1}^{C} \cdots \sum_{c_m=1}^{C} \kappa_{c_1} \cdots \kappa_{c_m} \prod_{i=1}^{n} \left[\sum_{r=1}^{R} \pi_r \prod_{j=1}^{m} f(y_{ij}|x_{ij}, \theta_{r_i c_j}) \right]. \quad (2)$$

Introducing a $C \times 1$ latent column group membership variable W_j (with $W_{jc} = 1$ if $j \in c$ and $W_{jc'} = 0$ for $c' \neq c$) alongside the latent row group membership variable Z_i the joint distribution of the augmented data is

$$f(y_{ij}, Z_i, W_j|x_{ij}, \Omega) = \prod_{r=1}^{R} \prod_{c=1}^{C} \left[\pi_r \kappa_c f(y_{ij}|x_{ij}, \theta_{rc})\right]^{Z_{ir} W_{jc}}$$

leading to the complete data likelihood

$$L_c(\Omega|\{y_{ij}\}, \{Z_{ir}\}, \{W_{jc}\}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \prod_{r=1}^{R} \prod_{c=1}^{C} \left[\pi_r \kappa_c f(y_{ij}|x_{ij}, \theta_{rc})\right]^{Z_{ir} W_{jc}}.$$

In the case of the student feedback form data set, biclustering implies the simultaneous clustering of students and questions into student clusters and question clusters.

2.3 Specific models

We now present specific expressions for the finite mixture model component distributions for binary, Poisson count and two specific ordinal data types. Generalisations to other count types (e.g. Negative Binomial) and other ordinal models are straightforward. The building blocks of the likelihood are the probability distributions

$$f(y|\theta) = \begin{cases} \theta^y (1-\theta)^{1-y} & \text{Binary} & y \in \{0, 1\} \\ e^{-\theta}\theta^y / y! & \text{Poisson count} & y \in 0, 1, 2, \ldots \\ \prod_{k=1}^{q} \theta_k^{I(y=k)} & \text{Ordinal} & y \in \{1, 2, \ldots, q\}. \end{cases} \quad (3)$$

In the ordinal case we have a variable with q levels $y \in \{1, \ldots, q\}$ and $\sum_{k=1}^{q} \theta_k = 1$.

In this paper we focus on models where the model parameter θ in (3) can be constructed from a linear predictor of the general form $\eta = \mu + x^T \beta$ for some parameter vector β and covariates x. For binary variables use the logit link

$$\eta = \text{logit}(\theta) = \text{logit}(P[Y = 1]) = \mu + x^T \beta$$

and use the log link for count variables

$$\eta = \log(\theta) = \log(E[Y]) = \mu + x^T \beta.$$

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

With ordinal variables we use one of two models. The proportional odds model has

$$\eta_k = \text{logit}\left(\sum_{\ell=1}^{k}\theta_\ell\right) = \text{logit}(P[Y \le k]) = \mu_k - x^T\beta \tag{4}$$

with $\mu_1 \le \mu_2 \le \ldots \le \mu_{q-1} \le \mu_q = +\infty$. The ordering of the μ_k parameters gives the model its ordinal character, and the negative sign in (4) is a convention that ensures that higher covariate values make higher values of Y more likely. An alternative ordinal model is the ordered stereotype model (Anderson 1984) which has

$$\eta_k = \log\left(\frac{\theta_k}{\theta_1}\right) = \log\left(\frac{P[Y=k]}{P[Y=1]}\right) = \mu_k + \phi_k x^T\beta$$

with score parameters $\phi_1 = 0 \le \phi_2 \le \ldots \le \phi_{q-1} \le \phi_q = 1$. These score parameters have the appealing interpretation as a numerical representation of the category levels, possibly unevenly spaced.

Clustering is introduced by having the linear predictor depend on the (unmeasured) latent row and/or cluster membership, as well as any measured covariates. Those covariates are now being absorbed into the set of parameters θ so that we add the row and column subscripts to θ_{ij} to reflect this in the following sections.

2.3.1 The row-clustered model

For row-clustered binary and count models the linear predictor for observation y_{ij} conditional on $i \in r$ is

$$\text{logit}(\theta_{ijr}) \ \text{ or } \ \log(\theta_{ijr}) = \eta_{ijr} = \mu + \alpha_r + \beta_j + \gamma_{rj} + x_{ij}^T\delta_{rj}$$

with $E[y_{ij}|x_{ij}, i \in r] = \theta_{ijr}$, and corner point or sum to zero identifiability constraints on $\{\alpha_r\}$, $\{\beta_j\}$ and $\{\gamma_{rj}\}$. The sets $\{\alpha_r\}$ and $\{\beta_j\}$ represent the parameters quantifying the main effects of the R row groups and m columns respectively, the set $\{\gamma_{rj}\}$ are the associations between the different row clusters and columns, and $\{\delta_{rj}\}$ represents the effects of the covariates. The additive version of these models omits the interaction term γ_{rj}. The two ordinal models have $P[y_{ij} = k|x_{ij}, i \in r] = \theta_{ijrk}$. The proportional odds ordinal model has

$$\text{logit}\left(\sum_{\ell=1}^{k}\theta_{ijr\ell}\right) = \eta_{ijrk} = \mu_k - \alpha_r - \beta_j - \gamma_{rj} - x_{ij}^T\delta_{rj}$$

and the ordered stereotype model has

$$\log\left(\frac{\theta_{ijrk}}{\theta_{ijr1}}\right) = \eta_{ijrk} = \mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj} + x_{ij}^T\delta_{rj}).$$

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

The complete data log likelihood of these models, using the known data $\{y_{ij}\}$ and the assumed latent class memberships $\{Z_{ir}\}$, is as follows

$$\ell_c(\Omega|\{y_{ij}\},\{Z_{ir}\}) = \sum_{i=1}^{n}\sum_{r=1}^{R} Z_{ir}\log(\pi_r) + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{r=1}^{R} D_1(y_{ij}, Z_{ir}, \theta_{ijr}), \qquad (5)$$

where

$$D_1(y_{ij}, Z_{ir}, \theta_{ijr}) = \begin{cases} Z_{ir}\{y_{ij}\log(\theta_{ijr}) + (1-y_{ij})\log(1-\theta_{ijr})\}, & \text{Binary} \\ Z_{ir}(y_{ij}\log(\theta_{ijr}) - \theta_{ijr}), & \text{Poisson count} \\ \sum_{k=1}^{q} Z_{ir} I(y_{ij}=k)\log\left(\theta_{ijrk}\right), & \text{Ordinal.} \end{cases}$$

2.3.2 The biclustered model

For biclustered data the equivalent expressions are

$$\text{logit}(\theta_{ijrc}) \;\text{ or }\; \log(\theta_{ijrc}) = \eta_{ijrc} = \mu + \alpha_r + \beta_c + \gamma_{rc} + x_{ij}^T\delta_{rc}$$

for binary and count data models, with $E[y_{ij}|x_{ij}, i \in r, j \in c] = \theta_{ijrc}$ and identifiability constraints on $\{\alpha_r\}$, $\{\beta_c\}$ and $\{\gamma_{rc}\}$. For the ordinal models $P[y_{ij} = k|x_{ij}, i \in r, j \in c] = \theta_{ijrck}$. In the proportional odds model we have

$$\text{logit}\left(\sum_{\ell=1}^{k}\theta_{ijrc\ell}\right) = \eta_{ijrck} = \mu_k - \alpha_r - \beta_c - \gamma_{rc} - x_{ij}^T\delta_{rc}$$

and for the ordered stereotype model

$$\log\left(\frac{\theta_{ijrck}}{\theta_{ijrc1}}\right) = \eta_{ijrck} = \mu_k + \phi_k(\alpha_r + \beta_c + \gamma_{rc} + x_{ij}^T\delta_{rc}).$$

Consequently, the complete data log likelihood of this model using the known data $\{y_{ij}\}$ and the row and column memberships $\{Z_{ir}\}$ and $\{W_{jc}\}$ is as follows:

$$\begin{aligned}\ell_c(\Omega \mid \{y_{ij}\},\{Z_{ir}\},\{W_{jc}\}) &= \sum_{i=1}^{n}\sum_{r=1}^{R} Z_{ir}\log(\pi_r) + \sum_{j=1}^{m}\sum_{c=1}^{C} W_{jc}\log(\kappa_c) \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{r=1}^{R}\sum_{c=1}^{C} D_2(y_{ij}, Z_{ir}, W_{jc}, \{\theta_{ijrc}\}) \qquad (6)\end{aligned}$$

where

$$D_2(y_{ij}, Z_{ir}, W_{jc}, \{\theta_{ijrc}\}) = \begin{cases} Z_{ir}W_{jc}\{y_{ij}\log(\theta_{ijrc}) + (1-y_{ij})\log(1-\theta_{ijrc})\}, & \text{Binary} \\ Z_{ir}W_{jc}(y_{ij}\log(\theta_{ijrc}) - \theta_{ijrc}), & \text{Poisson count} \\ \sum_{k=1}^{q} Z_{ir}W_{jc}I(y_{ij}=k)\log\left(\theta_{ijrck}\right), & \text{Ordinal} \end{cases}$$

3 Estimation and model selection

3.1 Maximum likelihood

All the models in this paper are likelihood-based and may be fitted by maximum likelihood, by direct maximisation of the likelihoods (1) and (2). This yields parameter estimates and their estimated asymptotic standard errors from the observed information matrix. Possible multimodality of the likelihood surface necessitates trying multiple starting points to avoid being locked into a local maximum.

The likelihoods (1) and (2) are however computationally expensive to evaluate, due to the need to sum over all possible allocations of observations to clusters. More rapid estimation is available through the EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997) with the missing data being the group membership of each row and/or column.

The EM algorithm uses the formulae for the log likelihood under complete knowledge, denoted by ℓ_c (see their expressions for row clustering and biclustering in (5) and (6), respectively), to produce the estimates in the E and M steps. The E step of the algorithm provides estimates of the posterior probabilities of allocations to clusters. Conditional on the data, the covariates, and the current parameter estimates $E[Z_{ir}] = \widehat{z}_{ir}$ is the posterior probability that $i \in r$, and for biclustering $E[W_{jc}] = \widehat{w}_{jc}$ is the posterior probability that $j \in c$. Note that $\forall i, \sum_{r=1}^{R} \widehat{z}_{ir} = 1$ and $\forall j, \sum_{c=1}^{C} \widehat{w}_{jc} = 1$. Given these estimates of the latent group memberships the M step of the EM algorithm maximises the appropriate complete data log likelihood, (5) or (6) to update the parameter estimates Ω.

The use of EM algorithm to estimate the model parameters is exemplified in Pledger and Arnold (2014) for the Bernoulli and Poisson distributions, in Fernández et al. (2016) for the ordered stereotype model, and in Matechou et al. (2011) for the propotional odds model. In the E-step of the EM algorithm for the biclustering model, the expected value of the product term $E[Z_{ir}W_{jc}|\{y_{ij}\}, \widehat{\Omega}]$ in (6) is approximated using the variational approximation $E[Z_{ir}W_{jc}|\{y_{ij}\}, \widehat{\Omega}] \simeq E[Z_{ir}|\{y_{ij}\}, \widehat{\Omega}]E[W_{jc}|\{y_{ij}\}, \widehat{\Omega}]$ employed by Govaert and Nadif (2005). To ensure that this approximation does not affect any final estimates, Fernández et al. (2016) use the resulting approximate MLEs from the EM algorithm as starting points to directly numerically maximise the incomplete data log likelihood (2). We also note that during the maximisation 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 $r = 1, \ldots, R-1$

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

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 during the maximisation, taking values over the whole real line.

Once the models are fitted, they may be compared by likelihood ratio tests (LRTs). A standard LRT may be successful when attempting to determine the need to include particular covariates in the model, and the presence of fixed column effects $\{\beta_j\}$ in row clustered models, or the interaction $\{\gamma_{rj}\}$ terms. However, there is a failure of necessary regularity conditions for LRTs if the comparison is between models with different numbers of clusters—when certain parameters (certain π_r and κ_c values) lie on the boundary of parameter space (Self and Liang 1987). In these cases we may use the theory in Self and Liang (1987) or randomisation tests (McLachlan 1987; Manly 2007; Gotelli and Graves 1996) to obtain the distribution of the test statistic under the null hypothesis. Estimation of standard errors are available using the curvature of the (incomplete data) log likelihood.

Information criteria, for example AIC (Akaike's Information Criterion) or its small-sample modification AICc (Akaike 1973; Burnham and Anderson 2002), provide an alternative means not only for choosing which covariates/effects to include but for comparing models of different dimension. The identification of the number of clusters is, of course, a key outcome of any cluster analysis and a number of approaches have been proposed to solve this problem [see e.g. McLachlan (1982), McLachlan and Basford (1988), Fraley and Raftery (2002), Sugar and James (2003), Raftery and Dean (2006), McCullagh and Yang (2008), Silvestre et al. (2014), Hasnat et al. (2015)]. There are a number of information criteria available, however the choice of the best criterion appears to be highly situation dependent, despite strong theoretical reasons for preferring one criterion over another (Schwarz 1978; Biernacki et al. 1998; McLachlan and Peel 2000).

As a specific example demonstrating the behaviour of these criteria, we carried out an extensive simulation study comparing the performance of eleven information criteria. Our particular interest was to determine how well they could identify the number of clusters in ordinal data using the proportional odds model (Matechou et al. 2011) and the ordered stereotype model (Fernández and Arnold 2016). The criteria were AIC, AIC_c, BIC, ICL-BIC, AIC_u, AIC3, CLC, CAIC, NEC, AWE and the $\mathcal{L}$ criterion. (Their definitions are given in Table S1 in "Supplementary Appendix S1".) We tested a range of sample sizes and included situations where the true cluster sizes differed strongly, as well as cases where clusters had very similar parameter values.

Overall, variants of AIC performed the best. For row-clustered ordered stereotype models, AIC correctly selected the number of row clusters in 93.8% of cases, followed by AIC_c (89.8%) and AIC_u (82.4%). Similar results were found in biclustered models. AIC_c and AIC_u also perform very well with percentages close to AIC: 85.6% and 84.2% respectively. BIC, which has a stronger model complexity penalty, underestimates the number of clusters (incorrectly selecting a smaller number of clusters in 56% and 63.2% of cases in row clustering and biclustering respectively).

In the case of proportional odds models, AIC3 has the best performance (selecting the correct model in 78% of cases), followed by BIC (75%), AIC, AIC_c, AIC_u, and CAIC (73%).

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

The other criteria (ICL-BIC, CLC, AWE and NEC) in both settings showed poor performance in selecting the correct number of clusters.

3.2 Bayesian approaches

Bayesian estimation provides a practical and tractable alternative to maximum likelihood estimation [see e.g. McLachlan and Peel (2000), Lee et al. (2008)]. An important advantage of Bayesian methods is that parameter estimation and model selection methodologies do not depend on the regularity conditions required by the LRT and which are violated in the fitting of finite mixtures, and can apply without modification to large and small samples. Additionally, Bayesian approaches incorporate prior knowledge regarding the parameters, and the results include the whole joint posterior distribution of the parameters (see a review of advantages in Wagenmakers et al. (2008, Chapter 9). Bayesian models are however often more computationally intensive (particularly where estimated by Markov Chain Monte Carlo, MCMC, methods), and have additional complexities such as label switching (see below).

A good introduction to Bayesian modeling of finite mixtures was given by Marin et al. (2005), Jasra et al. (2005) and and Marin and Robert (2007, Chapter 6) and Frühwirth-Schnatter (2006) gave a detailed review of Bayesian methods for finite mixtures. There are numerous examples of applications to continuous data (Richardson and Green 1997; Fraley and Raftery 2007; Stahl and Sallis 2012, e.g.). There is however a lack of development of a Bayesian inference approach with mixture models for ordinal data. Such models have additional complexities including the need for priors ensuring the ordering of parameters ($\{\mu_k\}$ in the proportional odds model, and $\{\phi_k\}$ in the ordered stereotype model).

Trans-dimensional implementations of MCMC provide a straightforward means of identifying the number of clusters. In particular, the reversible jump MCMC (RJMCMC) algorithm, introduced by Green (1995), has a sampler which jumps between parameter vectors with different numbers of components R. The RJMCMC approach is attractive because it solves the parameter estimation and dimension finding problems simultaneously. An alternative is the birth-and-death process (Stephens 2000a), whose mechanism has been shown to be essentially the same as RJMCMC algorithm (Cappé et al. 2003). Examples of the application of this algorithm in the context of mixture models is given, for instance, in Marrs (1998), Zhang et al. (2004), and Dellaportas and Papageorgiou (2006).

Using a trans-dimensional method the analyst can estimate the number of components by restricting attention to the model with the highest posterior probability. Alternatively, where the posterior distribution does not concentrate strongly on a single model with a fixed dimension, model-averaged estimates of the dimension-independent parameters can be calculated easily, incorporating this additional model uncertainty. Fernández and Arnold (2016) investigated the choice of the number of components most suitable for a given data set in the context of row clustering of ordinal data modelled by the ordered stereotype model (Fernández et al. 2016). This work compared two methodologies for selecting the best model: the first approach fits a separate model to the data for each possible number of clusters using the EM

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

algorithm (Sect. 3.1). Information criteria are then used to select the best model. The second approach uses a trans-dimensional Bayesian construction in which the parameters and the number of clusters are estimated simultaneously from their joint posterior distribution. The results described in their paper for the RJMCMC sampler are encouraging in its ability to select models correctly. An outline of the RJMCMC sampler for one-dimension clustering is given in "Supplementary Appendix S2". The use of likelihood maximization to evaluate information criteria such as the AIC is difficult when the likelihood surface is flat or contains long level ridges. A particular advantage of a Bayesian approach is that the estimation process is more stable in those cases.

In a mixture model the labels $\{1, \ldots, R\}$ are not identifiable and are arbitrary. For example, the row cluster mixture model $\widehat{\pi}_1 f(y|x, \widehat{\theta}_1) + \widehat{\pi}_2 f(y|x, \widehat{\theta}_2)$ has the same likelihood when we replace estimates $(\widehat{\pi}_1, \widehat{\pi}_2, \widehat{\theta}_1, \widehat{\theta}_2)$ with $(\widehat{\pi}_2, \widehat{\pi}_1, \widehat{\theta}_2, \widehat{\theta}_1)$. Therefore, we cannot uniquely identify $\widehat{\pi}_1 f(\widehat{\Theta}_1; \boldsymbol{Y})$ as the "first" component of the mixture, and in an MCMC sampler the properties of a mixture component may be swapped many times with other components—leading to what is known as the 'label switching' problem (Stephens 2000b; Jasra et al. 2005). This problem can be resolved by placing an identifiability constraint (IC) on the parameters defining the mixture components. For example, we can require that $\alpha_1 < \alpha_2 < \cdots < \alpha_R$. Attractive as they are, ICs can often impede chain mixing and make it harder for the MCMC sampler to converge. A common alternative is to have no IC, but to relabel the components of the mixture after the sampler has run. There are a number of variants of relabelling procedures (Celeux 1998; Stephens 2000b; Frühwirth-Schnatter 2001; Hurn et al. 2003; Marin and Robert 2007). In our work we adopt the method introduced by Stephens (2000b), which is outlined in "Supplementary Appendix S3".

4 Visualising fitted models

The use of finite mixture approaches performs a fuzzy assignment of rows and/or columns to clusters, and therefore, any visualisation tool should take into account any fuzziness in the cluster structure. In this section, we present graphic tools for ordinal and count data sets (Sects. 4.1, 4.2, respectively). Two visualisation tools that represent this fuzziness are presented, which are based on the membership posterior probabilities $\{\widehat{Z}_{ir}\}$ that row i is in cluster r once we have observed the data $\{y_{ij}\}$ (Sect. 4.1.1), and the distances among score parameters $\{\widehat{\phi}_k\}$ when ordinal data is used (Sect. 4.1.2). A new graphical tool for ordinal data based on mosaic plots is described in Sect. 4.1.3 (Fernández et al. 2014). Section 4.2 shows graphical displays which are analogues of various existing and commonly used techniques in multivariate analysis (Pledger and Arnold 2014).

4.1 Ordinal data

The data we used to illustrate the graphical tools for ordinal data is the student feedback form ordinal data set. We fitted a suite of clustering models including row (student) clustering, column (question) clustering and biclustering (student and question). For each model, the information criteria AIC, AIC_c, BIC and ICL-BIC were computed and

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

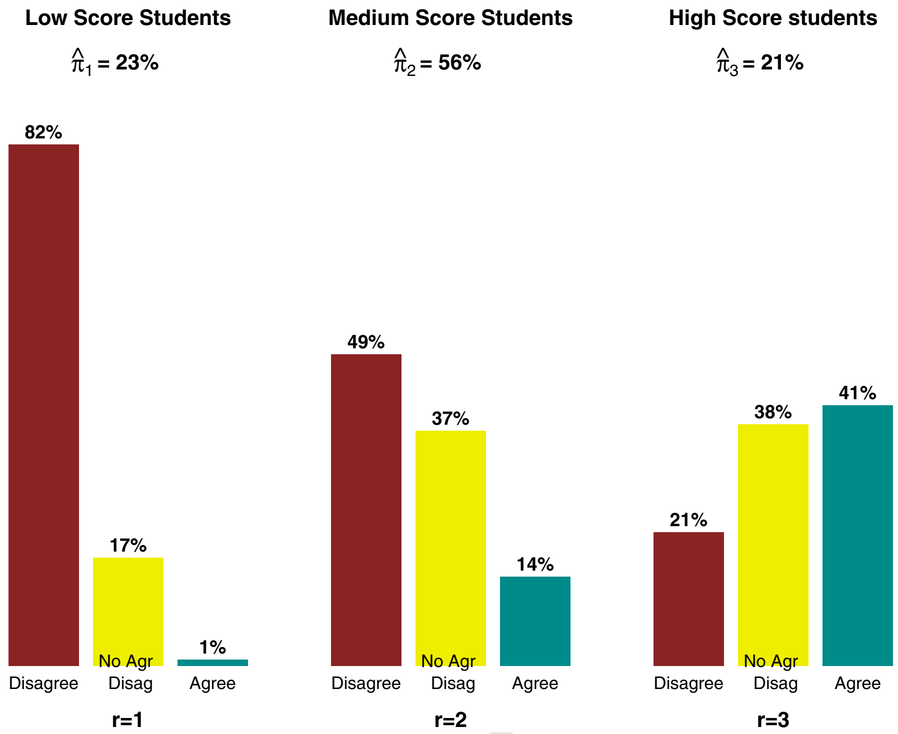

Fig. 1 $R = 3$ student group profiles. The percentage represents the estimated probability $\widehat{\theta}_{rk} = \sum_{j=1}^{m} \widehat{\theta}_{rjk}/m$ in each student group r and category k

the results are summarized in Table S6 in "Supplementary Appendix S5". Most of the information criteria indicate that the best clustering models are the ordered stereotype model version including row clustering with $R = 3$ row (student) groups and without interaction factors ($\mu_k + \phi_k(\alpha_r + \beta_j)$). Figure 1 displays the estimated probability $\widehat{\theta}_{rk} = \sum_{j=1}^{m} \widehat{\theta}_{rjk}/m$ of a member of group r responding at category level k. The students classified into the first group are those with lowest opinion of 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. More details about data set, list of questions, and traditional visualisation of the results (e.g. line plots and histograms) are given in Fernández et al. (2016).

4.1.1 Pairwise co-membership probabilities

Tibshirani and Walther (2005) developed a concept of strength of association based on the pairwise co-membership probabilities. The top graph in Fig. 2 shows a plot depicting the probability $C_{ii'}$ of any pair of students i and i' ($i, i' = 1, \ldots, n$) of being allocated to the same cluster for the data set with regard to students. The displayed probability $C_{ii'}$ in both contours is calculated as follows:

Journal: **11634** Article No.: **0324** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2018/5/7** Pages: **27** Layout: **Small-X**

$$
\begin{aligned}
C_{ii'} &= \sum_{r=1}^{R} P\left[Z_{ir}=1, Z_{i'r}=1 \mid \{y_{ij}\}, \widehat{\Omega}\right] \\
&= \sum_{r=1}^{R} P\left[Z_{ir}=1 \mid Z_{i'r}=1, \{y_{ij}\}, \widehat{\Omega}\right] P\left[Z_{i'r}=1 \mid \{y_{ij}\}, \widehat{\Omega}\right] \\
&= \sum_{r=1}^{R} P\left[Z_{ir}=1 \mid \{y_{ij}\}, \widehat{\Omega}\right] P\left[Z_{i'r}=1 \mid \{y_{ij}\}, \widehat{\Omega}\right] \\
&= \sum_{r=1}^{R} \widehat{Z}_{ir}\widehat{Z}_{i'r}, \qquad i, i' = 1, \ldots, n,
\end{aligned}
$$

where $\widehat{Z}_{ir}$ and $\widehat{Z}_{i'r}$ are the posterior probabilities that row i and i' respectively are members of row group r. It is important to note that we are assuming that the rows are independent conditional on the parameter vector Ω.

The contour plot is sorted by taking into account the column structure and the $R = 3$ clusters are clearly visible. Red tones represent pairs of students with a high probability of being allocated to the same cluster. Otherwise, orange tones are the students with a moderate probability and yellow tones are those students with lower probability of being allocated to the same cluster. Thus, this pairwise graph of the individuals can depict the cluster structure with the advantage of including the fuzzy assignment of rows to clusters based on the posterior probabilities $\{\widehat{Z}_{ir}\}$.

4.1.2 Fitted scores

For ordinal data, an alternative way of depicting the fuzziness of the probabilistic clustering is by means of the fitted score parameters from the ordinal stereotype model. The average fitted scores of each row (student) i across all of the m columns (questions) are:

$$
\overline{\phi}_{(i\cdot)} = \frac{1}{m} \sum_{j=1}^{m} \sum_{r=1}^{R} \sum_{k=1}^{q} \widehat{z}_{ir} \widehat{\phi}_k P[y_{ij} = k | i \in r] \qquad i = 1, \ldots, n.
$$

From here, we can compute the distance $D_{ii'} = |\bar{\phi}_{(i.)} - \bar{\phi}_{(i'.)}|$ based on the $\{\overline{\phi}_{(i\cdot)}\}$ values for any two rows (students) i and i' so that the differences between the fitted spacing of the levels of the ordinal response can be depicted. The full definition of the average score in the ordinal stereotype model is given in "Supplementary Appendix S4". The fuzziness in the clustering is shown in the bottom plots in Fig. 2 using a cell colour which goes from dark green to light brown. A dark green cell represents two students with a small distance in their fitted scores and who are therefore very likely to be in the same cluster. A light brown cell depicts high spacing distance between two students and a low probability of being in the same cluster. The rows were sorted according to the row cluster structure over both axes. As we noted on the fuzzy clustering heat maps (top graph), the three clusters are easily identifiable on the right level plot. The student cluster allocation is done by maximal posterior membership, i.e. each student

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

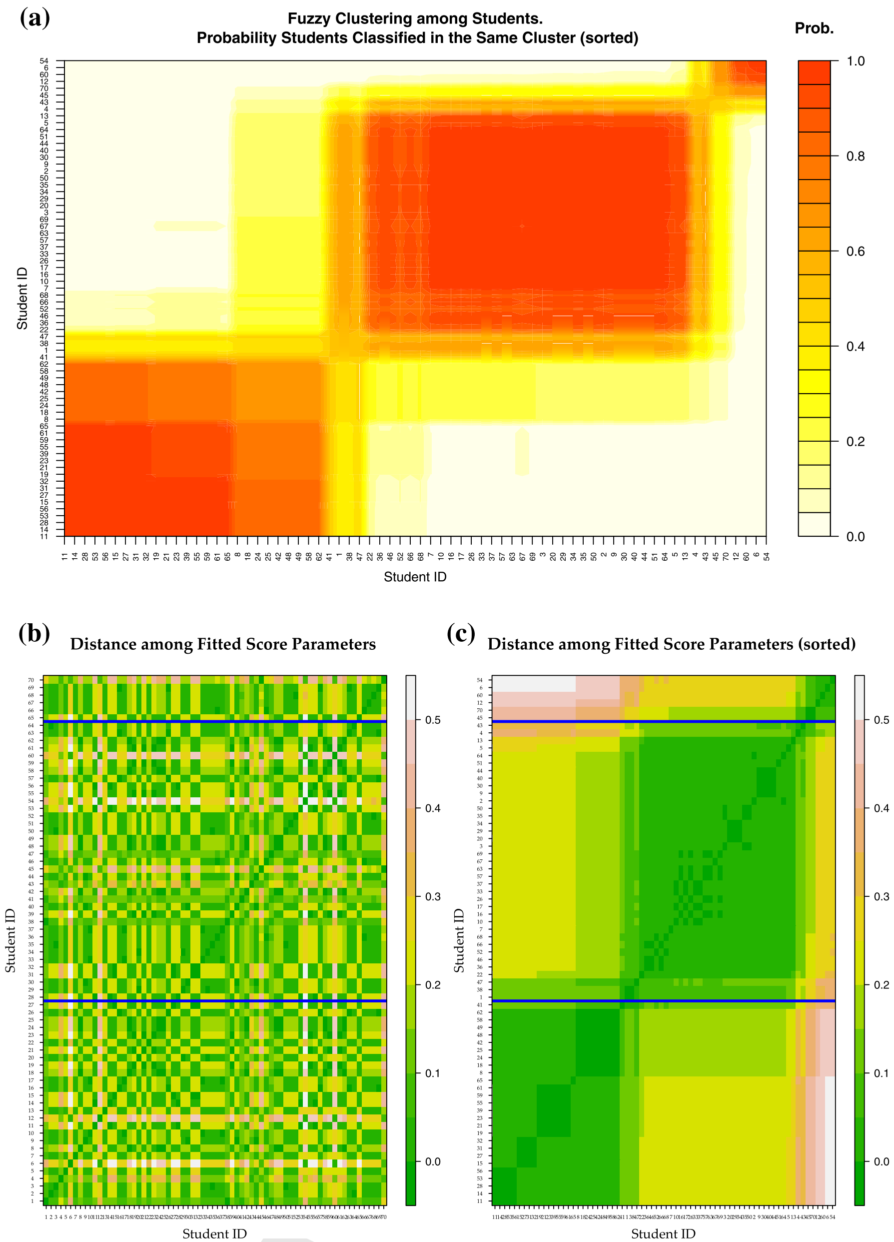

Fig. 2 Student feedback forms data set: the upper graph **a** shows a heat map of the pairwise probabilities that each student is a member of the same cluster. The students are sorted by the ($R = 3$) row cluster structure. The lower graph shows heat maps of the mean response level of each student to each question, (Eq. (S6) in "Supplementary Appendix S4"), with students and questions in (**b**) their original ordering and **c** ordered by cluster. The horizontal blue lines divide the plot to show the 3 clusters. The student cluster allocation is done by maximal posterior membership. The student orderings in (**a**) and (**c**) are the same

Journal: **11634** Article No.: **0324** TYPESET DISK LE CP Disp.:**2018/5/7** Pages: **27** Layout: **Small-X**

is allocated to the student group to which he or she belongs with the highest posterior probability. The student orderings in Fig. 2a, c are the same.

4.1.3 Spaced mosaic plots

Fernández et al. (2014) introduced a new graphical tool for ordinal data based on mosaic plots. The original mosaic plot was developed by Hartigan and Kleiner (1981) and refined by Friendly (1991). It is a graphical method for visualizing data from two qualitative variables which gives an overview of the data, makes it possible to recognize relationships, and shows the cross-sectional distribution of different variables. In this summary paper, we apply this visualization tool to the model-based methodology for matrices of ordinal data clustered using the ordered stereotype model. Therefore, the ordinal response level ($y \in \{1, \ldots, q\}$) and the cluster identity ($r \in \{1, \ldots, R\}$) in the data are considered as those two qualitative variables. Fernández et al. (2014) incorporated the estimated score parameters $\{\phi_k\}$ into the mosaic plot. As is mentioned in Sect. 2.3, those parameters determine the distance between two adjacent ordinal categories based on the data (see Anderson (1984); Agresti (2010) for more detail). For instance, in the student feedback form data set, the estimate of $\widehat{\phi_2}$ is 0.66. Therefore, given fixed values $\phi_1 = 0$ and $\phi_3 = 1$, it means that the space between "disagree" and "neither agree or disagree" is higher (0.66) than the space between "neither agree or disagree" and "agree" (0.34). The inclusion of space within a regular mosaic plot generates an enriched graph with more information which we called the *spaced mosaic plot*.

Figure 3 depicts a spaced mosaic plot of the student feedback forms data set for the model with row clustering with $R = 3$ student groups and $q = 3$ ordinal categories. The plot has three horizontal bands, one for each student cluster, with the height of each band proportional to the number of students in the cluster. Within each cluster, the vertical lines separate the ordinal responses, with the width of each block showing the proportions of responses in each category. Each block is labelled with the actual (relative) frequency. The blocks are held apart by rods representing the distances; in Fig. 3 the yellow rods are 0.66 units ($\widehat{\phi_2} - \widehat{\phi_1}$) and the red are 0.34 ($\widehat{\phi_3} - \widehat{\phi_2}$). Thus we can immediately see that categories 2 and 3 are close to each other, without needing to refer to the numerical values of $\widehat{\phi_k}$.

The *spaced mosaic plot* allows us to see at once the relative sizes of the row groups, the relative frequencies of the different response categories within each row group and the differences between the levels of the response categories. More details may be found in Fernández et al. (2014). The main features of the spaced mosaic plots for ordinal data and the R function to implement it are described in "Supplementary Appendix S6".

The construction of this new plot can be performed for one-dimensional clustering as shown, and also, by further subdividing the blocks, for biclustering. For instance, Fig. 4 shows a spaced mosaic plot with $R = 2$ student (row) clusters (y-axis) and $C = 3$ question (column) clusters (z-axis) for the ordinal student feedback form data set. The description of the graph is the same as explained for the one dimensional case. The only difference is that we use different colours to differentiate the column boxes

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

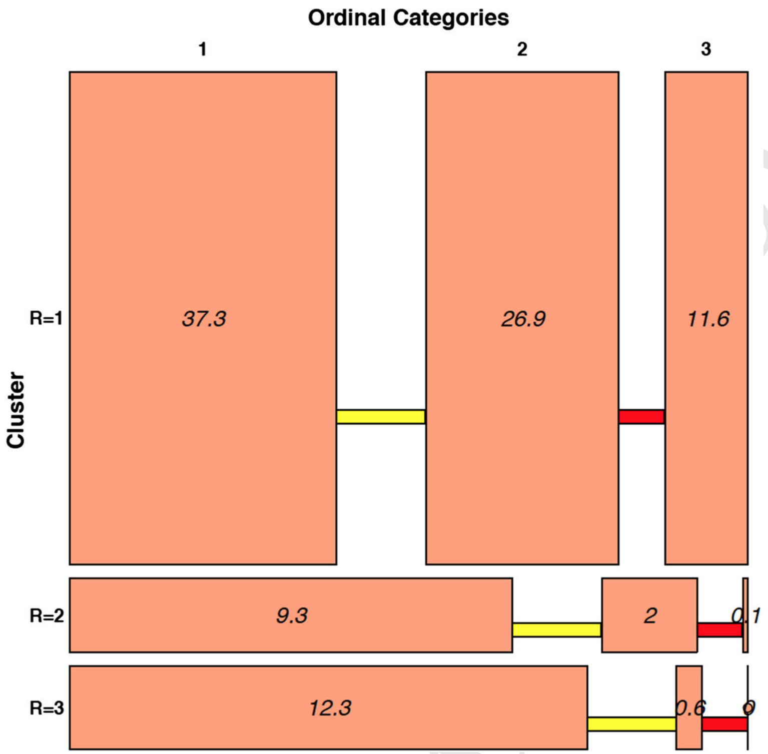

Fig. 3 Spaced mosaic plot for the row clustering model $R = 3$ for the student feedback forms data set. The height of each block is proportional to the number of rows in each row cluster; the width is proportional to the numbers of each ordinal responses within each row cluster. The area represents the frequency of each combination, also shown numerically in each block. The relative spacing between ordinal categories (e.g. 0.66 between 0 and 1, shown by the yellow bars) has been determined by the data

within each row box. In this case, blue boxes correspond to column cluster $C = 1$, orange ones to column cluster $C = 2$, and brown ones to column cluster $C = 3$.

4.2 Count data: generalisations of biplots

Clustering provides likelihood-based dimension reduction, leading to informative plots showing the main features of the data (Pledger and Arnold 2014). Clustering the rows of a data matrix yields a profile plot of row groups (labelled RG1, RG2, etc.) and a scatter plot of individual columns, and vice versa for column clustering with column groups labelled CG1, CG2, etc. After allowing for main effects, the interactions seen in the biclustering provide biplots, showing associations among rows, row groups, columns and column groups. The scatterplots are analogues of multidimensional scaling, and the biplots are analogues of correspondence analysis plots, but with a likelihood basis.

We use a test data set to illustrate the data visualisation for some of these graphs. The test data is an 8×10 matrix of counts where the rows and columns are labelled

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

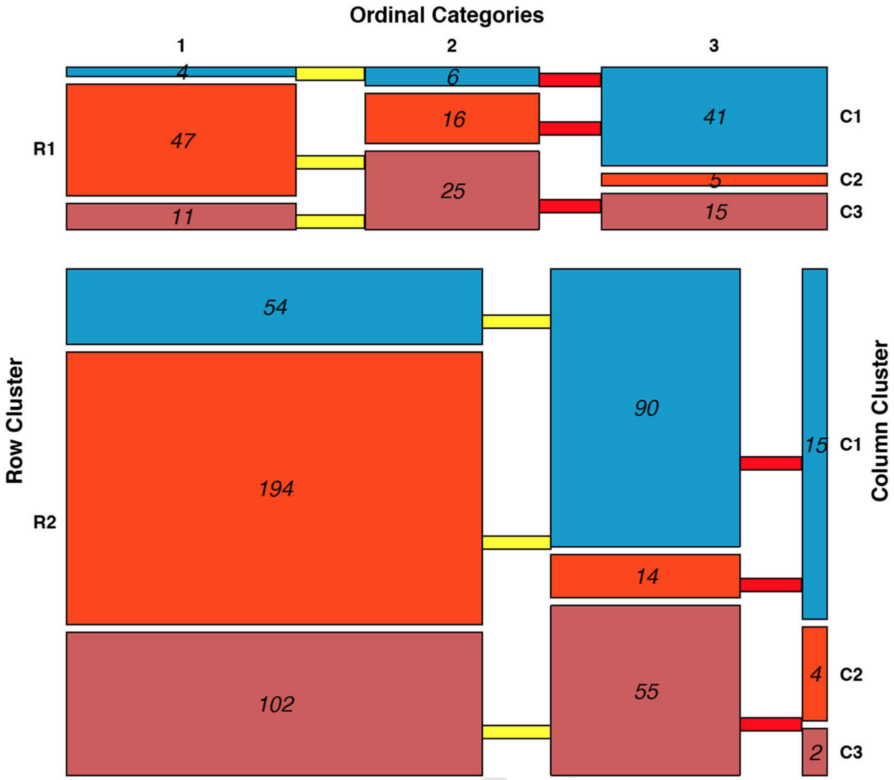

Fig. 4 Spaced mosaic plot for the student feedback forms data set for the biclustering model $R = 2$ student clusters and $C = 3$ question clusters

as $\{A, B, C, D, E, F, G, H\}$ and $\{a, b, c, d, e, f, g, h, i, j\}$, respectively. Figure 5 shows the test count data set.

For biclustering, a model with linear predictor

$$\mu + \alpha_i + \beta_j + \gamma_{rc}$$

adjusts for differing row and column sums (terms α_i and β_j respectively, the no-association model), allowing γ_{rc} to represent associations between row groups and column groups. For row clustering only, replace γ_{rc} with γ_{rj} to model associations between row groups and individual columns, and for column clustering only, use γ_{ic} to represent associations between individual rows and column clusters. In general the gamma values provide the plots in the link-transformed space, e.g. for row clustering each row r of γ_{rj} versus 1 to m shows the profile for row group r, while with R=3 the columns of γ_{rj} give coordinates in a plane embedded in 3-D space, thus providing a 2-D ordination diagram for the columns. However with a Poisson model special features of this distribution allow plotting in the original data space. The biplot methodology is to fit a 3 by 3 biclustering. The columns of γ_{rc} provide a scatterplot of the row groups, then imposing the same column clustering but allowing all rows to vary gives a matrix

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

Raw Count Data

	a	b	c	d	e	f	g	h	i	j
A	3	4	2	9	4	10	13	2	9	10
B	9	14	9	3	4	6	6	1	10	16
C	2	4	0	8	0	4	11	3	9	6
D	4	3	4	3	5	4	2	10	6	5
E	3	4	6	10	2	15	5	2	7	5
F	6	11	14	2	8	8	7	9	7	10
G	4	12	6	10	2	2	8	3	9	15
H	9	3	8	0	9	0	4	9	2	3

Fig. 5 Test data set: 8 × 10 matrix of counts where the rows and columns are labelled as $\{A, B, C, D, E, F, G, H\}$ and $\{a, b, c, d, e, f, g, h, i, j\}$, respectively

γ_{ic} which allows individual rows to be plotted on the same plane. Similarly the row clustering from the biclustered model provides a 2-D plot of the column clusters and the individual columns. From there standard biplot methodology allows these two planes to be superimposed to illustrate which rows and columns are similar to each other (Pledger and Arnold 2014).

The parameter γ is useful for displaying patterns in the data. For example with Poisson assumptions and row clustering into (say) three row clusters (groups, RG1, RG2 and RG3), the 3 by p table of estimates of (γ_{rj}) gives data for plotting three row-group profiles across all the different columns of the original data matrix (Fig. 6a). The same γ_{rj} table has three coordinates associated with each column of the original data, and hence provides a scatterplot of all the different data columns in 3 dimensions. However sum-to-zero constraints for the γ table ensure these points are coplanar (on triangle $A_1A_2A_3$ in Fig. 6b) and so may be rotated to be viewed more simply in two dimensions. Columns which are close in this scatterplot have similar data patterns. Similarly a model which clusters columns into three groups (CG1, CG2 and CG3) while keeping the rows separate provides an n by 3 table of pattern parameters (γ_{ic}). The columns of this table provide profiles of the three column groups over the different rows (Fig. 6c) while the rows of this pattern table give a scatterplot of the separate data rows in 3 dimensions (coplanar in triangle $B_1B_2B_3$ in Fig. 6d, and hence able to be rotated down into a simple 2-dimensional plot). A biclustering allows the two triangles to be rotated and superimposed (using a singular value decomposition, SVD) to give a biplot (Fig. 6e). This is an alternative to the traditional biplot from correspondence analysis (Fig. 6f). The difference between the methods is that with finite mixtures, likelihoods are used to reduce the dimensions, after which all components of the SVD are used in the biplot, whereas with correspondence analysis a full distance-based SVD is done and the dimension is then reduced, using the first two components to draw the biplot. Both types of biplot do dimension reduction and superposition of row

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

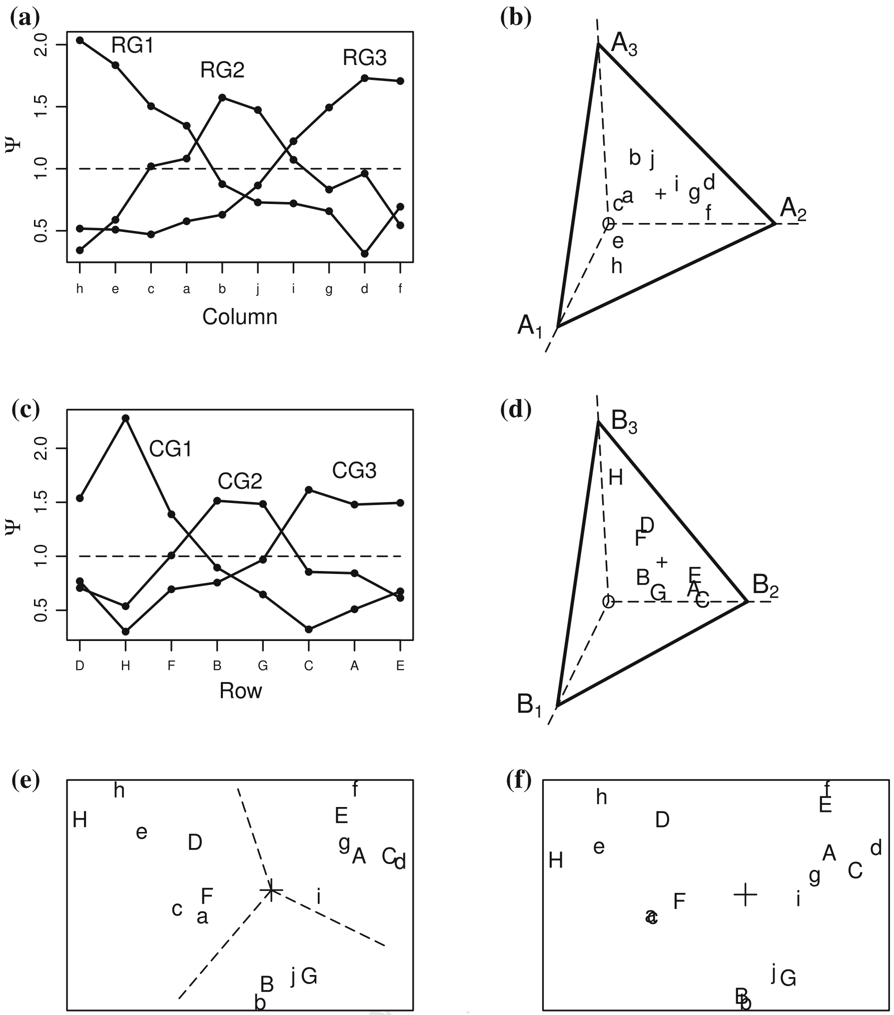

Fig. 6 Results of clustering a test data set into three row and three column groups. (See also Pledger and Arnold 2014). Plots **a**, **b** arise from row clustering and plots, **c**, **d** from column clustering. The biplot algorithm on **b** and **d** gives the combined plot **e**, which is similar to the standard correspondence analysis biplot in (**f**). Centroids are marked +

and column data; correspondence analysis uses mathematical distance measures while finite-mixture biclustering uses statistical likelihood measures.

5 Concluding remarks and extensions

This article summarises our recent contributions to mixture-based clustering and classification methods for binary, count and ordinal data. The common practice of treating ordinal data as continuous with equally spaced categories entails a loss of power and parsimony, and we have demonstrated a practical alternative in the clustering setting.

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

Perhaps the main challenge for the coming years is to make these methods better known to practitioners and researchers.

All the models are likelihood-based and may be fitted by maximum likelihood, yielding parameter estimates from the optimisation, and their estimated asymptotic standard errors from the observed information matrix. Maximum likelihood estimation provides advantages such as model comparison, hypothesis testing, and likelihood-based confidence intervals for parameters. Possible multimodality of the likelihood surface necessitates trying multiple starting points when using either direct optimisation or the EM algorithm to avoid being locked into a local maximum. We have had success using random starts combined with starting points found from using (double) k-means clustering (Maurizio 2001; Rocci and Vichi 2008). However it is almost impossible to provide general advice on the number of starting points required for all settings.

The models presented in this article may be also fitted with a Bayesian approach. A particular advantage of the trans-dimensional RJMCMC sampler, is the combination of the parameter estimation and model selection stages, and the computation of model specific and model averaged estimates are handled automatically. Alternatively, a single maximum a posteriori submodel can be selected if desired. Based on our experience, two of the drawbacks of the RJMCMC sampler are that it requires some care in the selection of suitable proposal distributions and the mixing can be slower than in fixed-dimensional MCMC samplers.

There are numerous applications for these models, for example in item response analysis and in contingency table analysis. The models presented here have been used for ecological (Pledger and Arnold 2014; Fernández et al. 2016; Fernández and Pledger 2016; Fernández and Arnold 2016), educational (Fernández et al. 2016), and medical (Matechou et al. 2011) applications to illustrate model fitting, fuzzy clustering, basic and pattern-detection models, binary, count and ordinal data, and the analogues of ordination, multidimensional scaling and correspondence analysis, with the substantial advantage of having a likelihood-based foundation. Our models are not, of course, limited to these fields.

For clustering purposes, there are typically two main approaches to the analysis of repeated measurements: subject-specific models and transitional models (Diggle et al. 2002; Vermunt and Hagenaars 2004; Agresti 2013). Subject-specific models, also known as conditional or random-effects models, describe effects at the individual or unit level and jointly model the response and individual random effects. In the case of model-based clustering, these random effects arise from a latent variable so that these models are also known as latent random effects models (Vermunt and Dijk 2001; Bartolucci et al. 2014). Vermunt and Dijk (2001) formulated a latent class regression model with class-specific coefficients, that is a finite mixture of random-intercepts and random-coefficients model. More recently, Bartolucci et al. (2014) presented a mixture of latent AR(1) processes with different correlation coefficients by cluster but the same variance. Their model also includes covariates and can handle longitudinal binary, categorical and ordinal data.

On the other hand, the transitional approach covers models in which past responses are included as predictors. These models are known as latent transition and Markov chain clustering models and typically use first-order Markov chains with states corresponding to the levels of the response. Frydman (2005), Pamminger and Frühwirth-Schnatter (2010), and Frühwirth-Schnatter et al. (2012) used this approach for model-based clustering of longitudinal categorical data. The latter two incorporate the effect of covariates in the cluster membership probabilities, use time-homogeneous Markov chains, and estimate their models within a Bayesian approach. Frydman (2005) considered a constrained version model where the transition matrices for the latent clusters are function of one of them. Estimation in this model is carried out using the EM algorithm. More recently, Costilla et al. (2015) proposed a Bayesian latent transitional approach for repeated ordinal data.

Data collection exercises commonly lead to data that are of mixed types: the data may be any of binary, nominal, ordinal, count or continuous variables. Multivariate analyses, in which multiple variables are treated simultaneously as outcomes, are typically restricted by the assumption that the data are all of a single type. However, there has thus far been little work on mixed type multivariate outcomes, despite the abundance of mixed type data sets. There has only been a small number of fully likelihood based treatments of the general multivariate mixed data problem where m variables of mixed types are measured on n individuals (Browne and McNicholas 2012; Cai et al. 2011; McParland and Gormley 2016). We are working on extending the likelihood based methods presented in this paper for finding association and correlation structures within potentially large multivariate data sets of mixed types.

In the analysis presented in this paper, we have considered only individuals with complete records, excluding participants with missing data. Missing data are often present in similar studies; and, hence, future work could extend the models to deal with such issues. Fitting the models using a Bayesian approach could provide a way of dealing with the missing data and also of choosing the right number of clusters, as, for example, in van Dijk et al. (2009) and Wyse and Friel (2012).

Another research direction would be to include the empirical study of models with interactions and the development of an extra layer in the RJMCMC sampler allowing both jumps between different class families (i.e., between models from the same family with and without interaction). We also envisage allowing jumps between one-dimensional (row or column clustering) and two-dimensional models (biclustering).

Fernández and Liu (2016) introduced a new goodness-of-test for ordered stereotype models based on the Hosmer–Lemeshow test for logistic regression and its version for the proportional odds model. A direct extension would be to develop a new goodness-of-fit measure which must take into account the possible clustering structure to reducing the dimensionality of the problem and become a parsimonious model. This new measure could be applied to all models presented in this article.

Acknowledgements This work was supported by the Marsden Fund on "Dimension reduction for mixed type multivariate data" (Award Number E2987-3648) from New Zealand Government funding, administrated by the Royal Society of New Zealand.

References

Agresti A (2010) Analysis of ordinal categorical data, 2nd edn. Wiley series in probability and statistics. Wiley, Hoboken

Agresti A (2013) Categorical data analysis, 3rd edn. Wiley series in probability and statistics. Wiley, Hoboken

Agresti A, Lang JB (1993) Quasi-symmetric latent class models, with application to rater agreement. Biometrics 49(1):131–139

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) 2nd international symposium on information theory, pp 267–281

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

Arnold R, Hayakawa Y, Yip P (2010) Capture-recapture estimation using finite mixtures of arbitrary dimension. Biometrics 66(2):644–655

Bartolucci F, Bacci S, Pennoni F (2014) Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. J R Stat Soc Ser C (Appl Stat) 63(2):267–288

Biernacki C, Celeux G, Govaert G (1998) Assessing a mixture model for clustering with the integrated completed likelihood. Technical Report 3521, INRIA, Rhne-Alpes

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

Breen R, Luijkx R (2010) Assessing proportionality in the proportional odds model for ordinal logistic regression. Sociol Methods Res 39(1):3–24

Browne RP, McNicholas PD (2012) Model-based clustering, classification, and discriminant analysis of data with mixed type. J Stat Plan Inference 142(11):2976–2984

Burnham KP, Anderson DR (2002) Model selection and multi-model inference: a practical information-theoretic approach, 2nd edn. Springer, Berlin

Cai JH, Song XY, Lam KH, Ip EHS (2011) A mixture of generalized latent variable models for mixed mode and heterogeneous data. Comput Stat Data Anal 55(11):2889–2907

Cappé O, Robert C, Rydén T (2003) Reversible jump, birth-and-death, and more general continuous time MCMC samplers. J R Stat Soc Ser B 65(3):679–700

Celeux G (1998) Bayesian inference for mixtures: the label switching problem. In: Proceedings in computational statistics 1998 (COMPSTAT98), Physica-Verlag HD, pp 227–232

Costilla R, Liu I, Arnold R (2015) A Bayesian model-based approach to estimate clusters in repeated ordinal data. In: JSM Proceedings, biometrics section, pp 545–556

Dellaportas P, Papageorgiou I (2006) Multivariate mixtures of normals with unknown number of components. Stat Comput 16(1):57–68

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

DeSantis SM, Houseman EA, Coull BA, Stemmer-Rachamimov A, Betensky RA (2008) A penalized latent class model for ordinal data. Biostatistics 9(2):249–262

Diggle PJ, Heagerty PJ, Liang KY, Zeger SL (2002) Analysis of longitudinal data second edition. Oxford statistical science series 1(25) [3]

van Dijk B, van Rosmalen J, Paap R (2009) A Bayesian approach to two-mode clustering. Technical Report

Everitt BS, Landau S, Leese M, Stahl D (2011) Cluster analysis, 5th edn. Wiley, Chichester

Fernández D, Arnold R (2016) Mode selection for mixture-based clustering for ordinal data. Aust NZ J Stat 58:437–472

Fernández D, Liu I (2016) A goodness-of-fit test for the ordered stereotype model. Stat Med 35(25):4660–4696

Fernández D, Pledger S (2016) Categorising count data into ordinal responses with application to ecological communities. J Agric Biol Environ Stat 21(2):348–362

Fernández D, Pledger S, Arnold R (2014) Introducing spaced mosaic plots. Research Report Series. ISSN: 1174-2011. 14-3, School of Mathematics, Statistics and Operations Research, VUW. http://msor.victoria.ac.nz/foswiki/pub/Main/ResearchReportSeries/TechReport_Spaced_Mosaic_Plots.pdf

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

Fraley C, Raftery AE (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J 41(8):578–588

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631

Fraley C, Raftery AE (2007) Bayesian regularization for normal mixture estimation and model-based clustering. J Classif 24(2):155–181

Friedman HP, Rubin J (1967) On some invariant criteria for grouping data. J Amer Stat Assoc 62:1159–1178

Friendly M (1991) Mosaic displays for multiway contingency tables. Technival Report 195, Department of Psychology Reports, New York University

Frühwirth-Schnatter S (2001) Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. J Am Stat Assoc 453(96):194–209

Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Wiley, New York

Frühwirth-Schnatter S, Pamminger C, Weber A, Winter-Ebmer R (2012) Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts markov chain clustering. J Appl Econom 27(7):1116–1137

Frydman H (2005) Estimation in the mixture of markov chains moving with different speeds. J Am Stat Assoc 100(471):1046–1053

Goodman LA (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61:215–231

Gotelli NJ, Graves GR (1996) Null models in ecology. Smithsonian Institution Press, Washington

Govaert G, Nadif M (2003) Clustering with block mixture models. Pattern Recognit 36(2):463–473

Govaert G, Nadif M (2005) An EM algorithm for the block mixture model. IEEE Trans Pattern Anal Mach Intell 27(4):643–647

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

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732

Haberman SJ (1979) Analysis of qualitative data, vol 2. Academic Press, New York

Hartigan JA, Kleiner B (1981) Mosaics for contingency tables. In: Proceedings of the 13th symposium on the interface between computer sciencies and statistics, Springer, pp 268–273

Hartigan JA, Wong MA (1979) Algorithm as 136: a k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat) 28(1):100–108

Hasnat MA, Velcin J, Bonnevay S, Jacques J (2015) Simultaneous clustering and model selection for multinomial distribution: a comparative study. In: International symposium on intelligent data analysis, Springer, pp 120–131

Hui FK, Taskinen S, Pledger S, Foster SD, Warton DI (2015) Model-based approaches to unconstrained ordination. Methods Ecol Evol 6(4):399–411

Hurn M, Justel A, Robert CP (2003) Estimating mixture of regressions. J Comput Graph Stat 12(1):55–79

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76(2):297–307

Jasra A, Holmes CC, Stephens DA (2005) MCMC and the label switching problem in Bayesian mixture models. Stat Sci 20(1):50–67

Jobson JD (1992) Applied multivariate data analysis: categorical and multivariate methods. Springer texts in statistics. Springer, Berlin

Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32(3):241–254

Lee K, Marin JM, Robert C, Mengersen K (2008) Bayesian inference on mixtures of distributions. In: Proceedings of the platinum jubilee of the Indian statistical institute, p 776

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Cam LML, Neyman J (eds) Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, University of California Press, pp 281–297

Manly BFJ (2005) Multivariate statistical methods: a primer. Chapman & Hall, London

Manly BFJ (2007) Randomization, bootstrap and monte carlo methods in biology, 3rd edn. Chapman & Hall, London

Marin JM, Robert C (2007) Bayesian core: a practical approach to computational Bayesian statistics. Springer texts in statistics. Springer, Berlin

Marin JM, Mengersen K, Robert C (2005) Bayesian modelling and inferences on mixtures of distributions. In: Dey D, Rao CR (eds) Handbook of statistics, vol 25. Springer, New York

Marrs AD (1998) An application of reversible-jump mcmc to multivariate spherical gaussian mixtures. In: Advances in neural information processing systems, pp 577–583

Matechou E, Liu I, Pledger S, Arnold R (2011) Biclustering models for ordinal data, presentation at the NZ Statistical Assn. In: Annual conference, University of Auckland, 28–31 Aug 2011

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

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

Maurizio V (2001) Double k-means clustering for simultaneous classification of objects and variables. In: Advances in classification and data analysis, pp 43–52

McCullagh P (1980) Regression models for ordinal data. J R Stat Soc 42(2):109–142

McCullagh P, Yang J (2008) How many clusters? Bayesian Anal 3(1):101–120

McCune B, Grace JB (2002) Analysis of ecological communities. Struct Equ Model 28(2)

McCutcheon AL (1987) Latent class analysis. Sage Publications, Thousand Oaks

McLachlan G, Peel D (2000) Finite mixture models. Wiley series in probability and statistics. Wiley, Hoboken

McLachlan GJ (1982) The classification and mixture maximum likelihood approaches to cluster analysis. Handb Stat 2(299):199–208

McLachlan GJ (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl Stat 36(3):318–324

McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Statistics, textbooks and monographs. M. Dekker, New York

McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley series in probability and statistics: applied probability and statistics. Wiley, Hoboken

McParland D, Gormley IC (2013) Clustering ordinal data via latent variable models. In: Lausen B, Van den Poel D, Ultsch A (eds) Algorithms from and for nature and life, studies in classification, data analysis, and knowledge organization. Springer, Berlin, pp 127–135

McParland D, Gormley IC (2016) Model based clustering for mixed data: clustMD. In: Advances in data analysis and classification, pp 1–15

Melnykov V (2013) Finite mixture modelling in mass spectrometry analysis. J R Stat Soc Ser C (Appl Stat) 62(4):573–592

Melnykov V, Maitra R (2010) Finite mixture models and model-based clustering. Stat Surv 4(9):80–116

Moustaki I (2000) A latent variable model for ordinal variables. Appl Psychol Meas 24(3):211–233

Nadif M, Govaert G (2005) A comparison between block CEM and two-way CEM algorithms to cluster a contingency table. In: European conference on principles of data mining and knowledge discovery, Springer, pp 609–616

Pamminger C, Frühwirth-Schnatter S et al (2010) Model-based clustering of categorical time series. Bayesian Anal 5(2):345–368

Pledger S (2000) Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics 56(2):434–442

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

Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge

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

Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components. J R Stat Soc Ser B 59(4):731–792

Rocci R, Vichi M (2008) Two-mode multi-partitioning. Comput Stat Data Anal 52(4):1984–2003

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

Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82(398):605–610

Silvestre C, Cardoso MG, Figueiredo MA (2014) Identifying the number of clusters in discrete mixture models. arXiv:14097419

Skrondal A, Rabe-Hesketh S (2004) Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. In: Monographs on statistics and applied probability. Chapman & Hall, London

Stahl D, Sallis H (2012) Model-based cluster analysis. Wiley Interdiscip Rev Comput Stat 4(4):341–358

Stephens M (2000a) Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. Ann Stat 28(1):40–74

Stephens M (2000b) Dealing with label switching in mixture models. J R Stat Soc Ser B 62(4):795–809

Sugar CA, James GM (2003) Finding the number of clusters in a dataset: an information-theoretic approach. J Am Stat Assoc 98(463):750–763

Tibshirani R, Walther G (2005) Cluster validation by prediction strength. J Comput Graph Stat 14(3):511–528

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X

Vermunt JK (2001) The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. Appl Psychol Meas 25(3):283–294

Vermunt JK, Hagenaars JA (2004) Ordinal longitudinal data analysis. In: Hauspie R, Cameron N, Molinari L (eds) Methods in human growth research. Cambridge University Press, Cambridge

Vermunt JK, Van Dijk L (2001) A nonparametric random-coefficients approach: the latent class regression model. Multilevel Model Newsl 13(2):6–13

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, Berlin, pp 43–52

Wagenmakers EJ, Lee M, Lodewyckx T, Iverson GJ (2008) Bayesian versus frequentist inference. Springer, Berlin

Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37

Wyse J, Friel N (2012) Block clustering with collapsed latent block models. Stat Comput 22(2):415–428

Zhang Z, Chan KL, Wu Y, Chen C (2004) Learning a multivariate gaussian mixture model with the reversible jump MCMC algorithm. Stat Comput 14(4):343–355

Journal: 11634 Article No.: 0324 TYPESET DISK LE CP Disp.:2018/5/7 Pages: 27 Layout: Small-X