Biclustering Models for Two-Mode Ordinal Data

Matechou, E, Liu, I, Fernandez, D, Farias, M & Gjelsvik, B

Original citation & hyperlink:
Matechou, E, Liu, I, Fernandez, D, Farias, M & Gjelsvik, B 2016, 'Biclustering Models for Two-Mode Ordinal Data' Psychometrika, vol 81, no. 3, pp. 611-624.
https://dx.doi.org/10.1007/s11336-016-9503-3

DOI 10.1007/s11336-016-9503-3
ISSN 0033-3123
ESSN 1860-0980

Publisher: Springer

The final publication is available at Springer via http://dx.doi.org/10.1007/s11336-016-9503-3

Biclustering models for ordinal data

Eleni Matechou*

Department of Statistics, University of Oxford, 1 South Parks Road, OX1 3TG, UK, +44(0)1865272593

Ivy Liu

School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, PO Box 600, Wellington 6140, NZ

Miguel Farias

Department of Experimental Psychology, University of Oxford, Tinbergen Building, 9 South Parks Road, Oxford, OX1 3UD, UK

Bergljot Gjelsvik

Oxford Mindfulness Centre, Department of Psychiatry, University of Oxford, Powic Building, Prince of Wales International Centre, Warneford Hospital, Oxford, OX3 7JX, UK

Abstract

The work in this paper introduces finite mixture models that can be used to simultaneously cluster the rows and columns of ordinal categorical response data. Model-fitting is performed using the EM algorithm and a fuzzy allocation of rows and columns to corresponding clusters is obtained. The clustering ability of the models is evaluated, and compared to that of k-means,

Email addresses: matechou@stats.ox.ac.uk (Eleni Matechou*), iliu@msor.vuw.ac.nz (Ivy Liu), miguel.farias@wolfson.ox.ac.uk (Miguel Farias), bergljot.gjelsvik@psych.ox.ac.uk (Bergljot Gjelsvik)

*Corresponding author

in a simulation study, and demonstrated using two real data sets.
Keywords: EM algorithm, fuzzy clustering, Likert scale, proportional odds.

1. Introduction

Measurement data with ordinal categories occur frequently and in many fields of application. For example in medicine, a continuous clinical response is often categorised into ordered subtypes based on histological or morphological terms. In a questionnaire, Likert scale responses might be "better", "unchanged" or "worse". Researchers often treat Likert scale responses as continuous. For example, the responses "better", "unchanged", and "worse" are assigned values "1", "2", and "3". The distance between the responses "better" and "unchanged" is then the same as the distance between "unchanged" and "worse". However, in reality, these distances might be different and unknown. In addition to the uncertainty of choices for the scores, Agresti (2010, Section 1.3) mentioned many other limitations, such as "ceiling effects" and "floor effects", of this naive approach, that can lead to misleading results. In this paper, we propose a biclustering method for analysing such data sets that uses only the ordering information.

The proportional odds model (McCullagh, 1980), which links the logits of cumulative probabilities with a set of predictors, is currently the most popular model for ordinal data while substantial developments in specialized methods for ordinal data have recently been made (see Liu and Agresti, 2005, for an overview). Nevertheless, there has been little work on cluster analysis for ordinal data. Traditional cluster analysis methods have been

used, but these wrongly treat the ordinal score as a continuous measurement and use matrix decomposition and eigenvalues for dimension reduction (Kaufman and Rousseeuw, 1990; Lewis et al., 2003). There also exist model-based approaches, however they do not fully incorporate the ordinal structure of the data in a probability model and instead use ad hoc distance metrics and crude similarity measures, such as Kendall's τ_b (Kendall, 1945), Goodman-Kruskal's γ (Goodman and Kruskal, 1954), and Somers' d (Somers, 1962). (see Podani, 2006; Giordan and Diana, 2011, for example).

Most model-based clustering methods related to finite mixtures deal with one-dimensional clustering, i.e. for the rows or for the columns only but not both, with the existing methods focusing on either continuous or categorical responses with no ordering nature (see Melnykov, 2013, for a recent review). Pledger and Arnold (2013) have recently developed a probabilistic model using finite mixtures that carries out a simultaneous fuzzy clustering of the rows and columns of binary or count data. The models, fitted using the Expectation-Maximisation algorithm (EM) (Dempster et al., 1977), provide a likelihood-based model analogue to multidimensional scale, correspondence analysis and association analysis.

In this paper, we generalise the Pledger and Arnold (2013) work to the case of ordinal categorical response data, and specifically using the proportional odds model parameterisation. The model structure is described in section 2. The quality of the clustering resulting from the model is evaluated, using simulation, is section 3.1. Finally, applications to two real data sets are shown in sections 3.2.1 and 3.2.2 and a discussion of the model and possible extensions is given in section 4.

2. Materials and Methods

2.1. Background: Proportional odds model

Consider the data set as an $n \times p$ matrix $\mathbf{Y}$ with entry y_{ij} the realisation of a multinomial distribution with q cells and $\theta_{ij1}, \ldots, \theta_{ijq}$ probabilities, $\sum_{k=1}^{q} \theta_{ijk} = 1, \forall i, j$. Let the set of model parameters be denoted by ϕ. The likelihood is formed as:

$$L(\phi|\mathbf{Y}) = \prod_{i=1}^{n} \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{ijk}^{I(y_{ij}=k)}, \tag{1}$$

where the indicator variable $I(\psi)$ is equal to 1 if condition ψ is satisfied and 0 otherwise.

Under the proportional odds model (McCullagh, 1980), and in the case where the additive effect of rows and columns on the response is considered:

$$\theta_{ijk} = \frac{\exp(\mu_k - \alpha_i - \beta_j)}{1 + \exp(\mu_k - \alpha_i - \beta_j)} - \frac{\exp(\mu_{k-1} - \alpha_i - \beta_j)}{1 + \exp(\mu_{k-1} - \alpha_i - \beta_j)}, \tag{2}$$

or alternatively,

$$\text{logit}\left[P(Y_{ij} \leq k)\right] = \mu_k - \alpha_i - \beta_j, \tag{3}$$

where μ_k is the k^{th} cut-off point, with $\mu_1 < \mu_2 < \ldots < \mu_{q-1}$, and α_i, β_j are respectively the effect of row i, column j on the response, with $\alpha_1 = \beta_1 = 0$. The total number of model parameters is equal to: $\nu = (q-1) + (n-1) + (p-1)$.

2.2. Biclustering: simultaneous clustering of rows and columns

Suppose that the rows come from a finite mixture with R components or row groups while the columns come from a finite mixture with C components

or column groups. If cell i, j belongs to row group r and column group c, then under the proportional odds model considered above:

$$\text{logit}\left[P(Y_{ij} \leq k)\right] = \mu_k - \alpha_r - \beta_c. \tag{4}$$

However, row and column group memberships are latent variables and therefore unknown. Define by Z_{ir} and X_{jc} the indicator random variables for group membership of row i in row group r and column j in column group c, respectively. The posterior probability that row i belongs to row group r is $E(Z_{ir}) = z_{ir},\ \sum_{r=1}^{R} z_{ir} = 1\ \forall i$, while the posterior probability that column j belongs to column group c is $E(X_{jc}) = x_{jc},\ \sum_{c=1}^{C} x_{jc} = 1\ \forall j$. $\widehat{z}_{ir}$ and $\widehat{x}_{jc}$ are obtained during the E-step of the EM algorithm:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \left\{ \sum_{c=1}^{C} \widehat{\kappa}_c \sum_{k=1}^{q} \widehat{\theta}_{rck}^{I(y_{ij}=k)} \right\}}{\sum_{a=1}^{R} \widehat{\pi}_a \prod_{j=1}^{p} \left\{ \sum_{b=1}^{C} \widehat{\kappa}_b \sum_{k=1}^{q} \widehat{\theta}_{abk}^{I(y_{ij}=k)} \right\}} \tag{5}$$

and

$$\widehat{x}_{jc} = \frac{\widehat{\kappa}_c \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \widehat{\pi}_r \sum_{k=1}^{q} \widehat{\theta}_{rck}^{I(y_{ij}=k)} \right\}}{\sum_{b=1}^{C} \widehat{\kappa}_b \prod_{i=1}^{n} \left\{ \sum_{a=1}^{R} \widehat{\pi}_a \sum_{k=1}^{q} \widehat{\theta}_{abk}^{I(y_{ij}=k)} \right\}}, \tag{6}$$

where $\widehat{\pi}_r$ and $\widehat{\kappa}_c$ correspond, respectively, to the estimated proportion of rows that belong to row group r and the estimated proportion of columns that belong to column group c, $\sum_{r=1}^{R} \widehat{\pi}_r = \sum_{c=1}^{C} \widehat{\kappa}_c = 1$. These proportions are estimated during the M-step of the algorithm by $\widehat{\pi}_r = \sum_{i=1}^{n} \widehat{z}_{ir}/n$ and $\widehat{\kappa}_c = \sum_{j=1}^{p} \widehat{x}_{jc}/n$.

The maximum likelihood estimates for the set of parameters ϕ, also obtained during the M-step of the algorithm, are found by maximising the log of the complete likelihood:

$$\ell_c(\phi|\mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{q}\widehat{z}_{ir}\widehat{x}_{jc}I(y_{ij}=k)\log(\theta_{rck})+\sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{z}_{ir}\log(\widehat{\pi}_r)+\sum_{j=1}^{p}\sum_{c=1}^{C}\widehat{x}_{jc}\log(\widehat{\kappa}_c). \tag{7}$$

The additive model shown in equation (4) can be extended to a model which allows for an interaction between the row and column effects by modelling the logits of the cumulative probabilities as:

$$\text{logit}\left[P(Y_{ij}\leq k)\right] = \mu_k - \alpha_r - \beta_c - \gamma_{r,c}. \tag{8}$$

Since $\sum_r \gamma_{r,c} = 0\ \forall c$ and $\sum_c \gamma_{r,c} = 0\ \forall r$, there are $(R-1)(C-1)$ more parameters introduced compared to the additive case.

The model can also be altered to consider one-dimensional clustering. The set of different models that can be fitted by considering row or column clustering alone, or both, with or without column/row effects, with or without interaction terms are shown in Table 1 with details given in Appendix A.

All the computer code, available here `http://www.stats.ox.ac.uk/~matechou/`, is written in **R** (**R** Development Core Team, 2010) and $\ell_c(\phi|\mathbf{Y})$ is maximised using the Newton-Raphson algorithm provided as an option in *optim*.

2.3. Model selection

Since these are likelihood-based models, likelihood-based model selection criteria, such as AIC (Akaike, 1973) or BIC (Schwarz, 1978), can be used to select amongst them. However, when comparing models with different numbers of clusters, the validity of these criteria is doubtful because of violation of regularity conditions (see McLachlan and Peel, 2000, section 6.4.2.). Despite the lack of theoretical foundations, the use of AIC and BIC has gained

Table 1: Model set with corresponding number of parameters ν. The following constraints are placed, where appropriate: $\alpha_1 = 0,\ \beta_1 = 0,\ \sum_r \gamma_{rj} = 0,\ \forall j,\ \sum_j \gamma_{rj} = 0,\ \forall r,\ \sum_{r=1}^{R} \pi_r = 1,\ \sum_{c=1}^{C} \kappa_c = 1$. $R = 1$: no row effect, $R = r$: r row groups and $R = n$: no row clustering. $C = 1$: no column effect, $C = c$: c column groups and $C = p$: no column clustering.

R	C	$\text{logit}\left[P(Y_{ij} \leq k)\right]$	ν
r	1	$\mu_k - \alpha_r$	$(q-1) + 2R - 2$
r	p	$\mu_k - \alpha_r - \beta_j$	$(q-1) + 2R + p - 3$
r	p	$\mu_k - \alpha_r - \beta_j - \gamma_{rj}$	$(q-1) + Rp + R - 2$
1	c	$\mu_k - \beta_c$	$(q-1) + 2C - 2$
n	c	$\mu_k - \alpha_i - \beta_c$	$(q-1) + 2C + n - 3$
n	c	$\mu_k - \alpha_i - \beta_c - \gamma_{ic}$	$(q-1) + Cn + C - 2$
r	c	$\mu_k - \alpha_r - \beta_c$	$(q-1) + 2R + 2C - 4$
r	c	$\mu_k - \alpha_r - \beta_c - \gamma_{rc}$	$(q-1) + RC + R + C - 3$

support in the literature although the first tends to overestimate the number of clusters needed (see Cubaynes et al., 2012; McLachlan and Peel, 2000, for simulation results and references).

Both criteria are of the form $-2\ell(\widehat{\phi}, \widehat{\pi}, \widehat{\kappa}|\mathbf{Y}) +$ a penalty term. $\ell(\widehat{\phi}, \widehat{\pi}, \widehat{\kappa}|\mathbf{Y})$ is the incomplete log-likelihood, which assuming row-based conditional independence in the biclustering case, is equal to:

$$\ell(\widehat{\phi}, \widehat{\pi}, \widehat{\kappa}|\mathbf{Y}) = \log \left[\sum_{c_1=1}^{C} \dots \sum_{c_p=1}^{C} \widehat{\kappa}_{c_1} \dots \widehat{\kappa}_{c_p} \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \widehat{\pi}_r \prod_{j=1}^{p} \prod_{k=1}^{q} \widehat{\theta}_{rc_jk}^{I(y_{ij}=k)} \right\} \right]. \quad (9)$$

The penalty term is equal to 2ν for AIC and $\log(np)\nu$ for BIC, where ν is the number of parameters in the incomplete likelihood i.e. the number of parameters in the complete likelihood plus $R + C - 2$ parameters for estimating the cluster proportions.

Evaluating the complete log-likelihood in reasonable computing time is feasible in the one-dimensional clustering case, but the same does not hold in the biclustering case, and especially for large data sets, as this evaluation requires consideration of all possible allocations of the p columns to the C groups, as expression (9) indicates.

An alternative criterion, the integrated classification criterion, is a classification-based information criterion developed by Biernacki et al. (2000) who showed that is has a similar behaviour to BIC and is easy to implement as it does not require evaluation of the incomplete log-likelihood. McLachlan and Peel (2000), who demonstrated that the integrated classification criterion correctly selected the true number of clusters in all 3 simulation sets that they considered, refer to it as ICL-BIC. It can be calculated as:

$$\text{ICL-BIC} = -2\ell_c(\widehat{\phi}|\mathbf{Y}) + \nu \log(np). \tag{10}$$

3. Results

3.1. Simulation study

Existing clustering methods treat ordinal categorical data as continuous. In this simulation study, we compare our proposed method, referred to as POFM for "proportional odds finite mixtures", with the k-means clustering (Hartigan and Wong, 1979), which is one of the most popular cluster analysis methods for continuous data. The k-means clustering can be used to partition the n rows into R clusters (or the p columns into C clusters). In the interest of comparing the performance of two methods, we consider clustering in one dimension of a data set of ordinal variables.

We simulated an underlying latent continuous measure y_{ij}^* from a logistic distribution with mean α_i^* and variance 1, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. To create the ordinal scale, we let $y_{ij} = 1$ if $y_{ij}^* \leq \mu_1$; $y_{ij} = k$ if $\mu_{k-1} < y_{ij}^* \leq \mu_k$ for all $k = 2, \ldots, q-1$; and otherwise, $y_{ij} = q$.

We set $R = 3$, i.e. considered the row-clustering case, and $\pi_1 = \pi_2 = \pi_3$ with all columns considered to be homogeneous. If two rows i and i' belong to the same cluster, say r, then $\alpha_i^* = \alpha_{i'}^* = \alpha_r$. We chose the cutpoints $\{\mu_k,\, k = 1, \ldots, q-1\}$ such that there is an equal probability of responding each of the response categories when a row falls in the first cluster. For example, when $q = 3$, $\mu_1 = \mu_2 = \log 2$. We varied n and p as $n = (9,\, 30,\, 99)$, $p = (10,\, 20,\, 100)$ and used $(\alpha_1,\, \alpha_2,\, \alpha_3) = (0,\, 1,\, 2)$, $(0,\, 2,\, 4)$, $(0,\, 1,\, 4)$, and $q = (3,\, 5,\, 7)$. When p is large, there are more data points for each row.

When q is large, the ordered categorical response has a finer scale. For the row cluster effects $\{\alpha_r,\, r = 1, 2, 3\}$, the last setting (0, 1, 4) gives an unbalanced effect where the difference between the first two clusters is small, but the first two clusters are quite different from the third cluster.

We used the pairwise information between rows to evaluate the performance of the two methods. For each simulated data set, we calculated the proportion of times that the pairwise grouping was correct across all ${}^nC_2$ pairs. Table 2 shows the average proportions of correct pairwise groupings for 1,000 simulated data sets for each of the scenarios.

All standard errors for the averages are less than 0.0026. Most of them are around 0.001. POFM performs better than k-means when the cluster effects are balanced. In general, the greater n, p, q, or the cluster effect are, the better the performance. The only case when k-means outperforms POFM is when $(\alpha_1,\, \alpha_2,\, \alpha_3) = (0,\, 1,\, 4)$. For this particular case, POFM failed to distinguish between clusters 1 and 2, and it partitioned the individuals into only two clusters, leaving one of the clusters empty. However, the quality of the row clustering is still satisfactory, with the mean proportions of correct pairwise groupings greater than 74% in all cases. The results naturally extend to the column clustering case.

3.2. Case-studies

3.2.1. Religious beliefs

The study, first published by Wiech et al. (2008), consisted of 24 individuals, the first 12 self-classified as religious and the remaining 12 self-classified as atheistic or agnostic. Participants replied to a questionnaire which included a Locus of Control scale (Levenson, 1981), which is widely used in

Table 2: The average proportion of times that the pairwise grouping is correct for 1,000 simulated data.

			$(\alpha_2, \alpha_3) = (1, 2)$			$(\alpha_2, \alpha_3) = (2, 4)$			$(\alpha_2, \alpha_3) = (1, 4)$		
n	p	method	$q = 3$	5	7	3	5	7	3	5	7
	10	POFM	0.61	0.63	0.64	0.73	0.78	0.80	0.74	0.75	0.75
		k-means	0.68	0.69	0.69	0.70	0.72	0.73	0.72	0.74	0.75
9	20	POFM	0.70	0.72	0.73	0.79	0.86	0.88	0.77	0.76	0.75
		k-means	0.70	0.71	0.72	0.71	0.73	0.74	0.74	0.77	0.78
	100	POFM	0.85	0.84	0.83	0.94	0.94	0.86	0.75	0.75	0.75
		k-means	0.74	0.77	0.78	0.74	0.77	0.78	0.79	0.88	0.90
	10	POFM	0.65	0.67	0.68	0.75	0.81	0.84	0.76	0.77	0.77
		k-means	0.66	0.67	0.68	0.70	0.72	0.73	0.71	0.74	0.76
30	20	POFM	0.73	0.76	0.77	0.84	0.93	0.95	0.78	0.78	0.78
		k-means	0.70	0.72	0.72	0.72	0.75	0.76	0.75	0.80	0.81
	100	POFM	0.94	0.92	0.91	0.95	0.99	0.92	0.77	0.77	0.77
		k-means	0.79	0.83	0.86	0.76	0.84	0.87	0.93	0.97	0.98
	10	POFM	0.67	0.68	0.69	0.76	0.84	0.88	0.76	0.77	0.78
		k-means	0.67	0.68	0.68	0.70	0.72	0.73	0.72	0.75	0.76
99	20	POFM	0.75	0.78	0.80	0.86	0.95	0.97	0.79	0.78	0.78
		k-means	0.71	0.73	0.74	0.73	0.77	0.80	0.79	0.85	0.86
	100	POFM	0.98	0.97	0.96	0.97	1.00	0.97	0.78	0.78	0.78
		k-means	0.88	0.92	0.93	0.82	0.87	0.89	0.99	0.99	0.99

Table 3: Models supported by ICL-BIC for the data set of example 3.2.1.

R	C	$\text{logit}\left[P(Y_{i,j} \leq k)\right]$	ν	ICL-BIC
24	3	$\mu_k - \alpha_i - \beta_c - \gamma_{i,c}$	78	2442.4
3	3	$\mu_k - \alpha_r - \beta_c - \gamma_{r,c}$	17	2446.6
24	2	$\mu_k - \alpha_i - \beta_j - \gamma_{i,c}$	53	2455.6

health and social psychology. The scale consisted of 32 questions, shown in Appendix B, all rated on a 6 - point Likert scale, (1)"Strongly disagree", ..., (6)"Strongly agree". The questions were designed to assess an individual's beliefs on the level of control that themselves - questions 1, 4, 5, 9, 21, 23, 27 and 29 - (internal control), powerful others - questions 3, 8, 11, 13, 16, 20, 25 and 30 - (external control), God - questions 15, 17, 19, 22, 24, 26, 28 and 31 - (God control) or luck and fate - questions 2, 6, 7, 10, 12, 14, 18 and 32 - (external control) had on their lives.

The models proposed in section 2 were fitted to the 24 by 32 matrix. Since the likelihood surface is multimodal, the EM algorithm is started from a number of different points to ensure that the best local maximum is obtained (Everitt et al., 2011). The model with the greatest support by ICL-BIC has $R = 24$, $C = 3$ and an interaction between row effects and column group effects. (Table 3).

The allocation of questions to column clusters is very clear, with all estimated posterior probabilities of allocation either practically 0 or 1. The three clusters separate the questions into three categories, internal control, God control and external control, respectively, almost perfectly. Cluster 1 includes questions 1, 4, 5, 9, 10, 23, 27 and 29, cluster 2 questions 15, 17, 19,

22, 24, 26, 28 and finally cluster 3 questions 2, 3, 6, 7, 8, 11, 12, 13, 14, 16, 18, 20, 21, 25, 30 and 32. This clustering generally replicates the factorial structure of the scale. Both luck/fate- and powerful others-related questions are underpinned by the belief that external factors command one's life, which makes it appropriate for them to be clustered together. Questions related to God and internal control formed two additional separate clusters, as the scale structure predicts.

The estimated probability of replying 3 or above to each of the 3 question clusters for all individuals is presented in Figure 1. All individuals tend to agree more with questions in column cluster 1 than cluster 3, regardless of their religious beliefs. However, there is a clear separation of the two groups in terms of column cluster 2.

3.2.2. Attempted suicides

The data set was collected as part of a study of patients admitted for deliberate self-harm (DSH) at the Acute Medical Departments of three major hospitals in Eastern Norway. We consider the answers of 151 individuals to 13 questions, shown in Appendix C, that were designed to assess the level of depression of the respondent by means of the Beck Depression Inventory-Short Form (BDI-SF) (Furlanetto et al., 2005). Response options range from 1 to 4, with higher scores indicating higher levels of depression (Beck et al., 1974).

We fitted biclustering models with $R = 2, \ldots, 5$ and C=2 or 3. The model supported by ICL-BIC has $R = 3, C = 2$ and an additive effect of row and column groups on the response. (Table 4).

The two column clusters are: (1, 2, 3, 4, 5, 7, 8, 10, 13) and (6, 9,

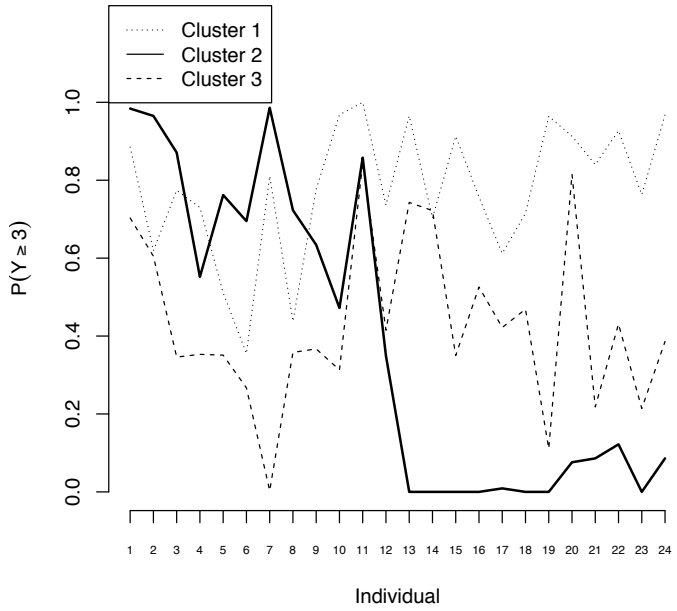

Figure 1: Estimated probabilities of replying 3 or above to each of the 3 column clusters for all 24 individuals, as derived by the selected model of Table 3.

11, 12), with the second cluster receiving lower scores than the first ($\widehat{\beta}_2 = -0.96(0.10)$), suggesting that these four questions are, possibly, markers of more severe forms of depression. The allocation of individuals to the 3 row groups, in proportions $0.282, 0.302, 0.416$, is relatively clear since all but 16 out of 151 individuals have one estimated posterior probability of allocation greater than 70%. The second row cluster is believed to show the most signs of depression since $\widehat{\alpha}_2 = 3.53(0.14)$ with the third cluster following with $\widehat{\alpha}_3 = 1.83(0.12)$. In fact, only 19.1% of individuals in cluster 2 contacted someone for help after their attempt, while the corresponding proportion

Table 4: Biclustering models supported by ICL-BIC for the data set of example 3.2.2.

R	C	$\text{logit}\left[P(Y_{i,j} \leq k)\right]$	ν	ICL-BIC
3	2	$\mu_k - \alpha_r - \beta_c$	9	4769.3
3	2	$\mu_k - \alpha_r - \beta_c - \gamma_{r,c}$	11	4784.2
4	2	$\mu_k - \alpha_r - \beta_c$	11	4794.3

for clusters 3 and 1 is 30% and 35.7%, respectively, which demonstrates the greater determination of individuals in cluster 2 to succeed in their attempt. Additionally, the proportion of individuals in clusters 2, 3 and 1 that had at least one episode of DSH within 3 months after the study is, respectively, equal to 27.8%, 15.9% and 9.5%. DSH is one of the most robust predictors of subsequent death by suicide (Hawton et al., 2013). The risk of suicide among DSH patients treated at hospital is 30- to 200-fold in the year following an episode compared to individuals with no history of DSH (Owens et al., 2002; Cooper et al., 2005; Hawton et al., 2012).

4. Discussion

Our biclustering models identify homogeneous groups of both rows and columns in data sets of ordinal responses, reducing the number of parameters needed to adequately describe the data and therefore easing interpretation. They fully account for the ordinal nature of the responses, while, being likelihood-based, give access to tools for selecting between possible models.

In the two real applications considered, both including questionnaire-type data designed to gain knowledge about the participants' personality, feelings and way of thinking, the clusters identified by the model agree with our

knowledge of the system and provide useful insight of the characteristics of the participants. Especially in the example of section 3.2.2 the way the participants were clustered agrees with information collected 3 months after the study was conducted.

In the analysis presented in section 3.2.2 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, or of appropriately averaging over models, using reversible jump MCMC (Green, 1995).

Although we have considered the proportional odds model parameterisation for the multinomial probabilities, the models extend easily to other ordinal models such as the adjacent-categories logit models, continuation-ratio logit models, and mean response models (see Agresti, 2012, for details on these models). Similarly, incorporating covariates to the model, when these are available, is straightforward by adjusting the linear predictor accordingly. The ordered stereotype model (Anderson, 1984), is another extension which we are currently considering.

We have presented the case when q, i.e. the number of levels, is the same for all variables. However, the models are easily extended to allow for a set of cut-points to be calculated for each unique value of q observed in the data set.

The area of application of these models is extremely wide and includes market research, where questions of the type "How likely are you to buy this

product in the future" have possible responses "Very likely to buy", "Likely to buy", "May or may not buy" etc. Additionally, the models are useful for services, such as websites, that review products, such as books, music albums, hotels etc. and provide recommendations to the users according to their own past reviews, as they can simultaneously cluster the individuals according to their taste, but also the products according to the reviews they have received from all users.

Future research will develop a graphical method for matrix visualisation, taking the resulting probabilities of allocation for each individual data point into account. The existing graphical methods rely on the use of ad hoc distance metrics and similarity measures which, as we have noted above, do not respect the full ordinal nature of the data.

Acknowledgements

We are grateful to Shirley Pledger for the discussions about the Pledger and Arnold (2013) paper.

Supplementary Materials

Appendices A, referenced in Section 2, B, referenced in Section 3.2.1, and C, referenced in Section 3.2.2, are available as supplementary material.

Agresti, A., 2010. Analysis of Ordinal Categorical Data, 2nd Edition. Wiley, New Jersey.

Agresti, A., 2012. Categorical data analysis. Wiley.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. B. N. Petrov, and F. Caski, (eds.) Proceeding of the Second International Symposium on Information Theory. Akademiai Kiado, Budapest., 267–281.

Anderson, J. A., 1984. Regression and ordered categorical variables. Journal of the Royal Statistical Society, Series B 46, 1–30.

Beck, A. T., Schuyler, D., Herman, I., 1974. Development of suicidal intent scales. In Beck, A.T., Resnik, H. L., Lettieri, D. J. (Eds) The prediction of suicide. Charles Press.

Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on pattern analysis and machine intelligence 22, No. 7.

Cooper, J., Kapur, N., Webb, R., Lawlor, M., Guthrie, E., Mackway-Jones, K., Appleby, L., 2005. Suicide after deliberate self-harm: a 4-year cohort study. American Journal of Psychiatry 162(2), 297–303.

Cubaynes, S., Lavergne, C., Marboutin, E., Gimenez, O., 2012. Assessing individual heterogeneity using model selection criteria: how many mixture components in capture-recapture models? Methods in Ecology and Evolution 3, 564–573.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

Everitt, B. S., Landau, S., Leese, M., Stahl, D., 2011. Cluster analysis. Wiley.

Furlanetto, L. M., Mendlowicz, M. V., Romildo Bueno, J., 2005. The validity of the Beck Depression Inventory-Short Form as a screening and diagnostic instrument for moderate and severe depression in medical inpatients. Journal of Affective Disorders 86, 87–91.

Giordan, M., Diana, G., 2011. A clustering method of categorical ordinal data. Communications in Statistics - Theory and Methods 40, 1315–1334.

Goodman, L. A., Kruskal, W. H., 1954. Measures of association for cross classifications. Journal of the American Statistical Association 49, 732–764.

Green, P. J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Hartigan, J. A., Wong, M. A., 1979. A k-means clustering algorithm. Applied Statistics 28, 100–108.

Hawton, K., Bergen, H., Kapur, N., Cooper, J., Steeg, S., Ness, J., Waters, K., 2012. Repetition of self-harm and suicide following self-harm in children and adolescents: findings from the Multicentre Study of Self-harm in England. Journal of Child Psychology and Psychiatry 53(12), 1212–1219.

Hawton, K., Casanas, I., Comabella, C., Haw, C., Saunders, K., 2013. Risk factors for suicide in individuals with depression: A systematic review. Journal of Affective Disorders 147(1-3), 17–28.

Kaufman, L., Rousseeuw, P. J., 1990. Finding groups in data: An introduction to cluster analysis. Wiley.

Kendall, M. G., 1945. The treatment of ties in ranking problems. Biometrika 33, 239–251.

Levenson, H., 1981. Differentiating among internality, powerful others, and chance. In H. M. Lefcourt (Ed.) Research with the locus of control construct. New York: Academic press 1, 15–63.

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. Journal of Neurology, Neurosurgery and Psychiatry 76, 343–348.

Liu, I., Agresti, A., 2005. The analysis of ordered categorical data: an overview and a survey of recent developments. Test 14, 1–73.

McCullagh, P., 1980. Regression models for ordinal data. Journal of the Royal Statistical Society. Series B. 42, 109–142.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley.

Melnykov, V., 2013. Challenges in model-based clustering. WIREs Computational Statistics 5, 135–148.

Owens, D., Horrocks, J., House, A., 2002. Fatal and non-fatal repetition of self-harm. Systematic review. Br J Psychiatry 181, 193–199.

Pledger, S., Arnold, R., 2013. Multivariate methods using mixtures: correspondence analysis, scaling and pattern detection. Computational Statistics and Data Analysis (online: http://dx.doi.org/10.1016/j.csda.2013.05.013).

Podani, J., 2006. Braun-Blanquet's legacy and data analysis in vegetation science. Journal of Vegetation Science 17, 113–117.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Somers, R. H., 1962. A new assymmetric measure of association for ordinal variables. American Sociological Review 27, 799–811.

Wiech, K., Farias, M., Kahane, G., Shackel, N., Tiede, W., Tracey, I., 2008. An fMRI study measuring analgesia enhanced by religion as a belief system. PAIN 139(2), 467–476.

Appendix A.

1. Row-clustering

$$\ell(\widehat{\phi}, \widehat{\pi}|\mathbf{Y}) = \sum_{i=1}^{n} \log \left(\sum_{r=1}^{R} \widehat{\pi}_r \prod_{j=1}^{p} \prod_{k=1}^{q} \widehat{\theta}_{rjk}^{I(y_{ij}=k)} \right)$$

E step:

$$\widehat{z}_{ir} = \frac{\widehat{\pi}_r \prod_{j=1}^{p} \prod_{k=1}^{q} \widehat{\theta}_{rjk}^{I(y_{ij}=k)}}{\sum_{a=1}^{R} \widehat{\pi}_a \prod_{j=1}^{p} \prod_{k=1}^{q} \widehat{\theta}_{ajk}^{I(y_{ij}=k)}}$$

M step: Numerically maximise:

$$\ell_c(\phi, \mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{k=1}^{q} \widehat{z}_{ir} I(y_{ij}=k)\log(\theta_{rjk}) + \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{z}_{ir}\log(\widehat{\pi}_r)$$

(a) No column effect (C=1):

$$\theta_{rjk} = \frac{\exp(\mu_k - \alpha_r)}{1+\exp(\mu_k-\alpha_r)} - \frac{\exp(\mu_{k-1}-\alpha_r)}{1+\exp(\mu_{k-1}-\alpha_r)}$$

(b) With column effect (C=p), no interaction:

$$\theta_{rjk} = \frac{\exp(\mu_k - \alpha_r - \beta_j)}{1+\exp(\mu_k-\alpha_r-\beta_j)} - \frac{\exp(\mu_{k-1}-\alpha_r-\beta_j)}{1+\exp(\mu_{k-1}-\alpha_r-\beta_j)}$$

(c) With column effect (C=p), interaction:

$$\theta_{rjk} = \frac{\exp(\mu_k - \alpha_r - \beta_j - \gamma_{rj})}{1+\exp(\mu_k-\alpha_r-\beta_j-\gamma_{rj})} - \frac{\exp(\mu_{k-1}-\alpha_r-\beta_j-\gamma_{rj})}{1+\exp(\mu_{k-1}-\alpha_r-\beta_j-\gamma_{rj})}$$

2. Column clustering

$$\ell(\widehat{\phi}, \widehat{\kappa}|\mathbf{Y}) = \sum_{j=1}^{p}\log\left(\sum_{c=1}^{C}\widehat{\kappa}_c\prod_{i=1}^{n}\prod_{k=1}^{q}\widehat{\theta}_{ick}^{I(y_{ij}=k)}\right)$$

E step:

$$\widehat{x}_{jc} = \frac{\widehat{\kappa}_c\prod_{i=1}^{n}\prod_{k=1}^{q}\widehat{\theta}_{ick}^{I(y_{ij}=k)}}{\sum_{a=1}^{C}\widehat{\kappa}_a\prod_{i=1}^{n}\prod_{k=1}^{q}\widehat{\theta}_{iak}^{I(y_{ij}=k)}}$$

M step: Numerically maximise:

$$\ell_c = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{c=1}^{C}\sum_{k=1}^{q}\widehat{x}_{jc}I(y_{ij}=k)\log(\theta_{ick}) + \sum_{j=1}^{p}\sum_{c=1}^{C}\widehat{x}_{jc}\log(\widehat{\kappa}_c)$$

(a) No row effect (R=1):

$$\theta_{ick} = \frac{\exp(\mu_k - \beta_c)}{1+\exp(\mu_k-\beta_c)} - \frac{\exp(\mu_{k-1}-\beta_c)}{1+\exp(\mu_{k-1}-\beta_c)}$$

(b) With row effect (R=n), no interaction :

$$\theta_{ick} = \frac{\exp(\mu_k - \alpha_i - \beta_c)}{1+\exp(\mu_k-\alpha_i-\beta_c)} - \frac{\exp(\mu_{k-1}-\alpha_i-\beta_c)}{1+\exp(\mu_{k-1}-\alpha_i-\beta_c)}$$

(c) With row effect (R=n), interaction:

$$\theta_{ick} = \frac{\exp(\mu_k - \alpha_i - \beta_c - \gamma_{ic})}{1+\exp(\mu_k-\alpha_i-\beta_c-\gamma_{ic})} - \frac{\exp(\mu_{k-1}-\alpha_i-\beta_c-\gamma_{ic})}{1+\exp(\mu_{k-1}-\alpha_i-\beta_c-\gamma_{ic})}$$

Appendix B.

1. Whether or not I get to be successful depends mostly on my ability. (Internal 1)
2. To a great extent my life is controlled by accidental happenings. (Fate 1)
3. I feel like what happens in my life is mostly determined by powerful people. (External 1)
4. Whether or not I get into a car accident depends mostly on how good a driver I am. (Internal 2)
5. When I make plans, I am almost certain to make them work. (Internal 3)
6. Often there is no chance of protecting my personal interests from bad luck happenings. (Fate 2)
7. When I get what I want, it is usually because I'm lucky. (Fate 3)
8. Although I might have good ability, I will not be given leadership responsibility without appealing to people in positions of power. (External 2)
9. How many friends I have depends on how nice a person I am. (Internal 4)
10. I have often found that what is going to happen will happen. (Fate 4)
11. My life is chiefly controlled by people who are more powerful than me. (External 3)
12. Whether or not I get into a car accident is mostly a matter of luck. (Fate 5)

13. People like myself have very little chance of protecting our personal interests when they conflict with those of strong pressure groups. (External 4)
14. It's not always wise for me to plan too far ahead because many things turn out to be a matter of good or bad fortune. (Fate 6)
15. What happens in my life is determined by God's purpose. (God 1)
16. Getting what I want requires pleasing those people above me. (External 5)
17. My life is primarily controlled by God. (God 2)
18. Whether or not I get to be successful depends on whether I'm lucky enough to be in the right place at the right time. (Fate 7)
19. When I am anxious, I rely on God for inner peace. (God 3)
20. If important people were to decide they didn't like me, I probably wouldn't make many friends. (External 6)
21. I can pretty much determine what will happen in my life. (Internal 5)
22. Whether or not I get into a car accident depends on God's plans. (God 4)
23. I am usually able to protect my personal interests. (Internal 6)
24. In order to have my plans work, I make sure they fit in with the commands of God. (God 5)
25. Whether or not I get into a car accident depends mostly on the other driver. (External 7)
26. When things don't go my way, I ought to pray. (God 6)
27. When I get what I want, it's usually because I worked hard for it. (Internal 7)

28. When faced with a difficult decision, I depend on God to guide my feelings and actions. (God 7)
29. My life is determined by my own actions. (Internal 8)
30. In order to have my plans work, I make sure they fit in with the desires of people who have power over me. (External 8)
31. When good things happen to me it is because of God's blessing. (God 8)
32. It's chiefly a matter of fate whether or not I have a few friends or many friends. (Fate 8)

Appendix C.

1. 1. I do not feel sad.
   2. I feel sad most of the time.
   3. I am sad all the time.
   4. I am so sad or unhappy that I can't stand it.
2. 1. I am not discouraged about my future.
   2. I feel more discouraged about my future than I used to do.
   3. I do not expect things to work out for me.
   4. I feel my future is hopeless and will only get worse.
3. 1. I do not feel like a failure.
   2. I have failed more than I should have.
   3. As I look back, I see a lot of failures.
   4. I feel I am a total failure as a person.
4. 1. I get as much pleasure as I ever did from the things I enjoy.
   2. I don't enjoy things as much as I used to.
   3. I get very little pleasure from the things I used to enjoy.
   4. I am so sad or unhappy that I can't stand it.
5. 1. I don't feel particularly guilty.
   2. I feel guilty over many things I have done or should have done.
   3. I feel quite guilty most of the time
   4. I feel guilty all of the time.
6. 1. I don't feel I am being punished.
   2. I feel I may be punished.
   3. I expect to be punished.
   4. I feel I am being punished.

7. 1. I feel the same about myself as ever.
   2. I have lost confidence in myself.
   3. I am disappointed in myself.
   4. I dislike myself.
8. 1. I don't criticise or blame myself more than usual .
   2. I am more critical of myself than I used to be
   3. I am sad all the time.
   4. I am so sad or unhappy that I can't stand it.
9. 1. I don't have any thoughts of killing myself.
   2. I have thoughts of killing myself, but I would not carry them out.
   3. I would like to kill myself.
   4. I would kill myself if I had the chance.
10. 1. I don't cry more than I used to.
    2. I cry more than I used to.
    3. I cry over every little thing.
    4. I feel like crying, but I can't.
11. 1. I am no more irritable than usual.
    2. I am more irritable than usual.
    3. I am much more irritable than usual.
    4. I am irritable all the time.
12. 1. I have not lost interest in other people.
    2. I am less interested in other people than before.
    3. I have lost most of my interest in other people.
    4. I have lost all my interest in other people.
13. 1. I make decisions about as well as ever.

2. I find it more difficult to make decisions than usual.
3. I have much greater difficulty in making decisions than I used to.
4. I have trouble making any decisions.