ONE-DIMENSIONAL MIXTURE-BASED CLUSTERING FOR ORDINAL RESPONSES

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

[1] University of Phayao, (e-mail: kemmawadee@gmail.com)

[2] Universitat Politècnica de Catalunya - BarcelonaTech, (e-mail: daniel.fernandez.martinez@upc.edu)

[3] Victoria University Wellington, (e-mail: ivy.liu@vuw.ac.nz, mcmilllo@ecs.vuw.ac.nz)

[4] University of Genoa, (e-mail: marta.nairuscone@unige.it)

[5] AgResearch NZ, (e-mail: roy.costilla@agresearch.co.nz)

ABSTRACT: Existing methods can perform likelihood-based clustering on a multivariate data matrix of ordinal responses, using finite mixtures to cluster the rows and columns of the matrix. Those models can incorporate the main effects of individual rows and columns and the cluster effects to model the matrix of responses. However, many real-world applications also include available covariates. In this study, we have extended mixture-based models to include covariates and test what effect this has on the resulting clustering structures. We focus on clustering the rows of the data matrix, using the proportional odds cumulative logit model for ordinal data. We fit the models using the Expectation-Maximization (EM) algorithm and assess their performance. Finally, we also illustrate an application of the models to the well-known arthritis clinical trial data set.

KEYWORDS: cluster analysis, mixture models , EM algorithm, ordinal responses, proportional odds mode.

1 Introduction

A well-known definition of an ordinal variable says it is one characterized by a categorical data scale, which describes an order showing differing degrees of dissimilarity (Agresti, 2010). Thus, although ordinal variables are affected by the distances among their ordinal categories, those distances are not known. In this work our approach to mixture-based clustering involves constructing an additive linear model of parameters, connected to the response data via a link function. Additional terms such as covariates may easily be added to the linear predictor. To the best of our knowledge, (Fernández *et al.*, 2019) introduced this formulation of model-based clustering for ordinal data with covariates, but the performance of these covariate methods and, more importantly, their influence on the resulting clustering structures, have not been documented so far. The main purpose of this article is to extend such models to include covariates and allow them to affect the detection of cluster structures. Moreover, we are also interested in comparing how the resulting clustering structures compare to those obtained without covariates, and how these changes may affect the interpretation of the results. We will focus on extending the one-dimensional clustering approach proposed in (Matechou *et al.*, 2016). This approach models ordinal response data using the proportional odds assumption of the cumulative logit model (from now on ”proportional odds model”). We will include covariates directly in the linear predictor.

2 Model formulation

When the data are in matrix form, clustering of rows is called row clustering. We present the row clustering formulation for finite mixtures based on the proportional odds model. This closely follows the model formulations in (Matechou *et al.*, 2016 , Fernández *et al.*, 2019). We decided to focus on row clustering because it is more common to have covariates linked to observations (rows) than to variables (columns). We consider a set of n subjects and m ordinal response variables, each with q possible ordinal response categories. Thus, data can be represented by an $n \times m$ matrix $\mathbf{Y}$ with ordinal entries y_{ij}. The row cluster index r ($r = 1,\ldots,R$) represents the number of the row cluster and the symbol $i \in r$ indicates that row i is allocated to row cluster r. We shall assume that all rows belonging to the same row cluster r have ordinal responses driven by the same row cluster effect, i.e. that there are no individual row effects. In a simpler model with clustering of rows, the rows (observations/subjects) will tend to be clustered if they have similar patterns of responses, without taking into account the information present in the covariates.

Having in mind that R and C are the numbers of row clusters and column clusters, respectively, we will deal with the possible values of $C = m$ (when column effects are different and therefore they are included within the model, without clustering). $C = 1$ when the column effect is the same and it is not included into the model.

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

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

where the parameters μ_k are the cutpoints and α_r indicates the effects of row cluster r. Adding p covariates into Model 1, we obtain

$$logit\left(\sum_{h=1}^{k}\theta_{ijrh}\right)=\eta_{ijrk}=\mu_k-(\alpha_r+x_i^T\delta_r), \tag{2}$$

where δ_r represent the effects of the covariates Models 1 and 2 will be used in the simulation and application section to compare the clustering structure.

3 Application

We applied the models proposed in this article to the *arthritis clinical trial* data set (Lipsitz *et al.*, 1996), which compares the drug auranofin and placebo therapy for the treatment of rheumatoid arthritis. The data set is obtained from the **R** package *multgee* (Touloumis, 2015). In this application, the covariate-dependent clustering could help to identify subsets of patients with similar covariate information patterns. This insight would be important because it would provide a flexible approach for identifying potential heterogeneous gender, age, and auranofin treatment effects on the arthritis scores. After fitting the models without covariates Eq.(1) and with covariates Eq.(2), with different number of row clusters, we compared them using the information criteria AIC and BIC (see results in Table 1). AIC indicates that the best model is the version of the row clustering model including age and treatment covariates ($\mu_k-(\alpha_r+x_{i1}\delta_{1r}+x_{i2}\delta_{2r})$) with $R=4$ row clusters (AIC = 2136.78), which is better than its counterpart in the model without covariates (AIC=2154.40). However, BIC shows that the model without covariates ($\mu_k-\alpha_r$) and $R=4$ is the best model (BIC=2202.05). A possible reason is that BIC penalizes higher numbers of parameters more strongly than AIC does, leading to a preference for more parsimonious models.

References

AGRESTI, ALAN. 2010. *Analysis of Ordinal Categorical Data, Second Edition*. Wiley Series in Probability and Statistics: John Wiley and Sons, Inc.

Table 1. *Results of row clustering models fitted to the arthritis data set. The best model in each group of models (no covariates, one, two, or three covariates), based on AIC, is shown in bold.*

Model		R	number of parameter	Log-like	AIC	BIC
$\mu_k - \alpha_r$		2	6	-1096.99	2205.99	2234.58
		3	8	-1077.73	2171.46	2209.59
		4	10	-1067.20	**2154.40**	**2202.05**
		5	12	-1067.20	2158.40	2215.58
$\mu_k - (\alpha_r + x_i\delta_r)$	x= age	2	8	-1138.18	2292.37	2330.49
		3	11	-1071.88	2165.75	2218.17
		4	14	-1065.18	2158.37	2225.08
		5	17	-1060.84	**2155.68**	2236.68
	x=treatment	2	8	-1082.28	2180.57	2218.69
		3	11	-1067.93	2157.87	2210.28
		4	14	-1057.70	**2143.40**	2210.11
		5	17	-1056.23	2146.46	2227.46
	x= gender	2	8	-1096.89	2209.77	2247.89
		3	11	-1079.51	2181.02	2233.44
		4	14	-1066.92	**2161.84**	2228.55
		5	17	-1066.37	2166.74	2247.74
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$	x_1 = age,	2	10	-1072.54	2165.07	2212.72
	x_2= treatment	3	14	-1059.23	2146.46	2213.17
		4	18	-1050.39	**2136.78**	2222.55
		5	22	-1048.53	2141.05	2245.88
	x_1 = age,	2	10	-1085.83	2191.67	2239.32
	x_2= gender	3	14	-1068.97	2165.95	2232.66
		4	18	-1061.29	**2158.58**	2244.35
		5	22	-1059.26	2162.52	2267.35
	x_1 = treatment,	2	10	-1081.82	2183.64	2231.29
	x_2= gender	3	14	-1065.99	2159.99	2226.71
		4	18	-1056.73	**2149.45**	2235.22
		5	22	-1055.06	2154.13	2258.96
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r} + x_{i3}\delta_{3r})$	x_1 = age,	2	12	-1071.60	2167.21	2224.39
	x_2= treatment,	3	17	-1060.50	2155.00	2236.01
	x_3= gender	4	22	-1050.35	**2144.71**	2249.54
		5	27	-1052.14	2158.35	2287.00

Fernández, Daniel, Arnold, Richard, Pledger, Shirley, Liu, Ivy, & Costilla, Roy. 2019. Finite mixture biclustering of discrete type multivariate data. *Advances in Data Analysis and Classification*, **13**, 117–143.

Lipsitz, Stuart R., Fitzmaurice, Garrett M., & Molenberghs, Geert. 1996. Goodness-of-Fit Tests for Ordinal Response Regression Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **45**(2), 175–190.

Matechou, Eleni, Liu, Ivy, Fernández, Daniel, Farias, Miguel, & Gjelsvik, Bergljot. 2016. Biclustering Models for Two-Mode Ordinal Data. *Psychometrika*, **81**(3), 611–624.

Touloumis, Anestis. 2015. R Package multgee: A Generalized Estimating Equations Solver for Multinomial Responses. *Journal of Statistical Software*, **64**(8), 1–14.