VICTORIA UNIVERSITY OF WELLINGTON

Te Whare Wānanga o te Ūpoko o te Ika a Māui

School of Mathematics and Statistics

Te Kura Mātai Tatauranga

Finite mixture model-based approach for clustering linguistics data

Minh Chau Van Nguyen

Supervisor: Louise McMillan

February 24th, 2020

STAT487 Research Paper

Abstract

This report documents the cluster analysis of ordinal and binary data using linguistics data from the Ewave data source. The model-based clustering approach using finite mixtures was proposed and the general ordinal models were described in the context of one-mode hard clustering and two-mode hard clustering. More specifically, row clustering and column clustering are one-mode clustering and biclustering is two-mode clustering. We introduced the proportional odds and the ordered stereotype models as the general ordinal models. These models with clustering were then fitted for each possible number of cluster using the Expectation-Maximisation (EM) algorithm. Model comparison based on information criteria is required for selecting the valid models with the correct number of clusters following model fitting and estimation. Visualization tools such as tables and graphs were used to interpret the clustering results of the Ewave data. We further set up a simulation study to test how reliably the EM algorithm is in a diverse range of scenarios (i.e. different combinations of starting points) in fitting the models and estimating the model parameters. Finally, considered related future works were highlighted.

Keywords— *cluster analysis, model-based clustering, finite mixture modelling, proportional odd model, ordered stereotype model, one-mode clustering, two-mode clustering, EM algorithm, simulation study, Information criteria (CRI)*

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr.Louise Mcmillan whom I respect so much for her vast knowlegde of just about everything, her extremely kind and patient personality and her non-stop encouragement. I am also very thankful for Louise's academic support and guidance as she has always been there to make it easier for me to put my project together. It has been an amazing experience working with my supervisor, given that I was always looking forward to going to our meetings and seeing her emails in my inbox. I genuinely look up to Louise to become a better statistician and most importantly a helpful person to others.

I would like to thank Professor Miriam Meyerhoff, Victoria University of Wellington for providing the data.

I would like to thank Dr.Ivy Liu for providing me all the possible reading materials there are and giving the permission to use the linguistics data sets I needed for writing this project.

I would like to make a special reference to Dr.Richard Arnold for explaining the data thoroughly in the beginning of the project, and for always being available when I had questions regarding the data set.

I would also like to thank Dr.Daniel Fernández for permission to use the spaced.mosaic.plot R function, and I would also like to thank Dr.Louise Mcmillan, Dr.Eleni Matechou and Dr.Daniel Fernández for permission to use the R clustord package.

1 Introduction

One-mode and two-mode clustering are the task of grouping data from a population into clusters based on their similar traits so that observations in the same cluster are similar and observations in separate clusters are different. For example, patients (rows) with heart disease can be clustered into one group basing on their heart disease related symptoms (columns), or the heart disease related symptoms (columns) can be clustered into one group basing on the patients (rows). These one-mode clustering types are known as row and column clustering, respective. Another type of clustering is the two-mode clustering, which refers to biclustering, where both the rows and columns are clustered simultaneously, e.g clustering in the above example group observations using both patients (rows) and their heart disease related symptoms (columns). The nature of clustering involves finding out which possible groups exist, as well as determining the unknown group memberships, i.e. we do not know which groups the individuals belong to in advance.

Broadly speaking, clustering can be divided into two subgroups: hard clustering where each data point either belongs to a cluster completely or not, and soft clustering where the clusters can overlap. This paper's purpose is to investigate and perform hard clustering of ordinal and binary data using a finite mixture model-based approach. In doing so, the goas are to find the true number of clusters in the data and to find which rows and/or columns are grouped together. Linguistic ordinal and binary data are used as real-life examples for data application to demonstrate how the model-based approach using finite mixture method can be applied to clustering.

Model-based clustering assumes there exists a model that is used to generate the data and that the true model from the data needs to be recovered. The recovered true model then defines the cluster structures by assigning rows and/or columns into clusters. When using a model-based approach, model selection and evaluation are then based on statistical inference methods, which allow choosing a suitable number of clusters of the data in a probabilistic framework. Finite mixture modelling method is a common method of model-based approach where each group (or cluster) is modelled by its own probability distribution and the overall distribution is a linear combination of these component distributions. This work details the definitions of finite mixture models used to analyze ordinal and binary data. The specific models considered for modelling ordinal data are the proportional odds model and the ordered stereotype model. In the clustering context, these models can be reformulated depending on the allocation of rows and/or column into clusters.

We perform cluster analysis by fitting the finite mixture models with clustering using the expectation-maximisation (EM) algorithm. The EM algorithm is a maximum likelihood estimation method that performs permutation of rows and/or columns to clusters based on the posterior probabilities. For better results, a simulation study is proposed to test how reliable the algorithm is in estimating the model parameters by running the algorithm multiple times, each time using a different combination of starting points.

The task of model selection is important and necessary for reliable and reproducible statistical inference. Thus, model selection is performed following model fitting for selecting the models with the true number of row and/or column clusters from a set of candidate models. A suitable and common approach for selecting the models is to use information criteria for comparing the candidate models. In particular, we compare how well the proportional odds model and the stereotype model fit the ordinal data; and compare how well the ordinal models and the binary models fit the same data coded as ordinal and binary.

This paper is organised into the following sections, **Section 2** is the literature review of the clustering concept and methods. **Section 3** provides the background of finite mixture models with clustering. **Section 4** outlines model fitting and parameters estimation using the iterative EM algorithm. Model selection is summarized in **Section 5**. **Section 6** consists of data application and data visualization. **Section 7** details the simulation study concerning the performance of the EM algorithm. The paper concludes with **Section 8**, a discussion.

2 Literature Review

The goal of cluster analysis is to find meaningful groups in data, allowing for pattern-detection and trend analysis. The results in addition consist not only of a partition of the data, but also of a characterization of the groups (Fruhwirth-Schnatter et al., 2019).

In general there are non-model-based and model-based approaches for cluster analysis (Fernández et al., 2018). Model-based clustering operates on the assumption that data originates from a finite mixture of underlying probability distributions (Agresti, 2010) where each cluster corresponds to a different component in the mixture. The traditional non-model-based approach on the other hand is heuristic and and not based on formal models (Agresti, 2010). Although non-model based methods are user-friendly in solving practical problems (Fernández et al., 2018), often the models proposed are not likelihood-based and hence statistical inference is not applicable (Pledger and Arnold, 2014). In contrast, model-based clustering via finite mixture not only allows the use of statistical inference and information criteria (Fernández et al., 2018) but it is also well-known for the better handling of missing data (Fernández et al., 2018). The latter advantage allows the finite mixture modelling method to fit well to the nature of clustering where the class membership is unknown (Dempster et al., 1977; McLachlan and Krishnan, 2007). This research explores row clustering, column clustering and biclustering of ordinal and binary data, using a model-based approach via finite mixtures modelling.

The linguistic data used throughout this paper is from the **Ewave** data source (Kortmann and Kerstin Lunkenheimer, 2013). The Ewave data presents grammatical variation in spontaneous spoken English, mapping 235 features in 76 varieties of English, where the response variable is described by the frequency in which each of the features can cross with the relevant variety of English (Kortmann and Kerstin Lunkenheimer, 2013). Ewave data was then processed and recoded as ordinal and binary data (Liu et al., 2019).

In using the model-based approach to model the linguistic data, the research goal could be to group the features in order to identify similarities and differences in grammatical structures between and within the features of different English languages. The finite mixture models used for clustering the data are the proportional odds model (Matechou et al., 2016; Costilla, 2017; Agresti, 2010), ordered stereotype model (Fernández, 2015; Fernández et al., 2016) and the binary model (Pledger and Arnold, 2014).

The clustering of ordinal data is common but doesn't always make the most of their ordinal nature (Costilla, 2017) as many statistical methods for categorical data treat all response variables like nominal, despite ordinal data having more information than nominal data (Costilla, 2017). Hence depending on the effect of the ordered characteristic, it could fail to report to the research goal. Moreover, using an ordinal analysis not only can provide relatively different and much powerful results than an analysis that ignores the ordering, but it also produces parsimonious models and hence simpler interpretation for the results (Agresti, 2010).

Previous cluster analyses of ordinal data typically used latent class models. By definition, the latent class models assume that there is some underlying variable called the latent variable (Fruhwirth-Schnatter et al., 2019) that influences what category each data point will take. This research however focuses on the use of the general ordinal models (Fernández et al., 2018), such as the proportional odds model and the ordered stereotype model. An important advantage of the proportional odds model is the model's parsimony of having few parameters required for describing the effect of the predictors on the ordinal response (Matechou et al., 2016). Additionally, the model parameters are invariant to the choice of categories for the ordinal response variable (Fernández et al., 2018). This ordinal model however has a drawback, where increasing the number of rows and/or columns in the data set will increase the number of parameters in the model thus interpretation for large data sets can be complex (Matechou et al., 2016). The main distinction between the proportional odds model and the stereotype model is that the latter model include the $\{\phi_k\}$ parameter, which can be interpreted as the scores of different categories of the response variable (Fernández and Pledger, 2015). These parameters then allow the stereotype model to be more flexible in comparison to the proportional odds model. In spite of this advantage, it is not as popular as the proportional odds model because the parameters are more difficult to estimate due to the intrinsic non-linearity combination of parameters in the predictors (Fernández et al., 2019). For binary data, General Linear Models (GLMs) are replaced by finite mixture models to model the data for there may be redundancy, with groups of associated rows (or columns) having similar incidence over the columns (or rows) (Pledger and Arnold, 2014). In general, it is desirable to find the most parsimonious models possible by applying both one-mode and two-mode clustering through finite mixtures with the aim to reduce the parameters of the saturated model.

The usual numerical approach to maximum likelihood estimation of finite mixture model parameters employs the procedure of the expectation-maximisation (EM) method. This algorithm is known to be widely used in hidden data problems, which fits perfectly onto the conditions of the mixture model of modelling missing data (Dempster et al., 1977; McLachlan and Krishnan, 2007) and hence applied well to the nature of clustering, where the missing information is the unknown cluster membership of each row and/or column (Fruhwirth-Schnatter et al., 2019). Many advantages come with the EM algorithm include providing a more stable way to convergence, and showing how the observations are assigned to their groups (Dempster et al., 1977; McLachlan and Krishnan, 2007). However, mixture modelling using EM algorithm can result in multimodality of the likelihood surface (Pledger and Arnold, 2014; Costilla, 2017). In relation to this issue, the EM algorithm may find local maximum likelihood estimates instead of the global maximum likelihood estimates (Dempster et al., 1977; McLachlan and Krishnan, 2007). Since the likelihood is multimodal, the EM algorithm is started from a number of different points and the iteration with the highest likelihood value is obtained (Matechou et al., 2016). Alternatively, Bayesian estimation using a Markov chain Monte Carlo (MCMC) procedure provides a practical and tractable estimation method. The benefits of a Bayesian approach may include better small sample results and greater flexibility in model selection (Fruhwirth-Schnatter et al., 2019). However, the Bayesian mixture modelling approach also has drawbacks to overcome such as the high computational cost in MCMC implementation to converge to the results, and the problem of label switching during MCMC sampling creating a lack of indentifiability (Fernández and Pledger, 2015).

The standard assumption of finite mixtures to clustering approach is that each of the components of the mixture model corresponds to a cluster solution. This implies that the component distribution specified in the mixture model is also the assumed cluster distribution (Fruhwirth-

Schnatter et al., 2019). As a result, the conclusions acquired from the estimated mixture model may be inaccurate when the estimated number of clusters is incorrect (Fernández, 2015). Therefore it is important to choose the valid model with the true number of clusters.

The fitted models may be compared using likelihood ratio tests (LRTs) and information criterion (CRI). LRTs are usually successful when attempting to determine whether to include predictors in the model, and the presence of fixed cluster effects in one-mode clustering models (Fernández et al., 2018). However, failure of necessary regularity conditions for LRTs is unavoidable if comparison is between models of different dimensions that have significant parameters lie on the boundary of parameter space (Fernández et al., 2018). Alternatively, likelihood-based methods such as AIC (Akaike's Information Criterion), its small-sample modification AIC$_c$, BIC (Bayesian information criterion) and its approximation ICL are proposed to compare between models of different structures, e.g with or without interactions, as well as to compare models of different number of clusters (Fernández et al., 2018).

3 The models

3.1 Ewave Data set

The Ewave data set can be described using the following tables. **Table 3.1** details the distribution of the 76 varieties of English in eight world regions. **Table 3.2** outlines the original rating of the grammatical features and **Table 3.3** outlines the recoding of the features rating after processing the data.

76 Varieties of English distribution			
British Isles **(11)**	Africa **(17)**	South Atlantic **(3)**	America **(10)**
Caribbean **(13)**	Southeast Asia **(8)**	Australia (5)	Pacific **(8)**

Table 3.1: World regions

235 grammatical features	
Code	**Description**
A	feature is pervasive or obligatory
B	feature is neither pervasive nor extremely rare
C	feature exists, but is extremely rare
D	attested absence of feature
X	feature is not applicable (given the structural make-up of the variety/P/C)
?	no information on feature is available

Table 3.2: Features ranking (Pre-processed data)

In terms of the data structure, the two-dimensional Ewave data can be represented by a $(n \times m)$ matrix $Y = \{y_{ij}\}$ where:

- n rows represent the varieties of English in the study, $\boldsymbol{i = 1, \dots, n}$
- m columns are the 235 features of the response variable y_{ij}, $\boldsymbol{j = 1, \dots, m}$
- k denotes the rating of the m features, $\boldsymbol{k = 1, 2, 3, 4}$ for $\boldsymbol{q = 4}$ ordinal categories for the ordinal data and $\boldsymbol{k = 1, 0}$ for the binary data.

235 grammatical features		
	Code	Description
Ordinal	1	present
	2	common
	3	rare
	4	absent
	NA	missing information
Binary	1	present
	0	absent

Table 3.3: Features ranking (Pre-processed data)

3.2 Background: Finite Mixture Models

Finite mixture modelling can be viewed as a latent variable analysis with a latent categorical variable describing the unknown group memberships corresponding to the different components of the mixture density (Fernández, 2015; Fernández et al., 2018). This approach assumes the data come from a mixture of specified number of groups G, where each observation is a realization y from the following finite mixture distribution (Fruhwirth-Schnatter et al., 2019):

$$y \sim \sum_{g=1}^{G} \eta_g f_g(y;\theta_g)$$

Here θ_g is the vector of unknown parameters in the gth component density f_g and $\eta_1, ..., \eta_G$ are mixing proportions which represent the a priori probability of being a member of group g, with constraint:

$$\sum_{g=1}^{G} \eta_g = 1, \qquad 0 \leq \eta_g \leq 1, \qquad g = 1, ..., G.$$

Suppose that the rows come from a finite mixture with R components or row clusters while the columns come from a finite mixture with C components or column clusters. Rows that belong to the same row cluster r are assumed to have the same effect on the response and columns that belong to the same column cluster c have the same effect on the response. The proportion of rows in row group r is π_r, and the proportion of columns in column group c is κ_c, with $\sum_{r=1}^{R} \pi_r = \sum_{c=1}^{C} \kappa_c = 1$. The finite mixture models with clustering are then formulated as follows,

$$y \sim \begin{cases} \bullet \ \sum_{r=1}^{R} \pi_r f_r(y_i;\theta_r), & \text{row clustered case,} \\ \bullet \ \sum_{c=1}^{C} \kappa_c f_c(y_j;\theta_c), & \text{column clustered case,} \\ \bullet \ \sum_{r=1}^{R} \pi_r \sum_{c=1}^{C} \kappa_c f_r c(y_{ij};\theta_{rc}), & \text{biclustered case,} \end{cases}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, m$, $r = 1, \ldots, R$, $c = 1, \ldots, C$.

And the component density f_g for the ordinal and binary data can be expressed as:

$$f(y,\theta) = \begin{cases} \bullet \ \theta^y (1-\theta)^{1-y}, & \text{Binary,} \quad y \in \{0,1\} \\ \bullet \ \prod_{k=1}^{q} \theta_k^{I(y=k)}, & \text{Ordinal,} \quad y \in \{1,2,\ldots,q\} \end{cases}$$

In the ordinal case we have a variable with q categories $y \in \{1, 2, \ldots, q\}$ and $\sum_{k=1}^{q} \theta_k = 1$. The binary and ordinal models are formed using parameter θ constructed from a linear predictor of the general form $\psi = \mu + \boldsymbol{\beta'x}$ for some parameter vector $\boldsymbol{\beta}$ and covariates $\boldsymbol{x}$. The details of these models including clustering by having the linear predictor depend on the unobserved latent row and/or cluster membership as well as the covariates, are described in the next section.

3.3 The proportional odds version of cumulative logit model

The proportional odds version of the cumulative logit model is also known as the logistic latent variable model, i.e. it assumes the ordinal response has an underlying continuous variable called a latent variable that follows a logistic distribution which has the same effects for each cumulative probability. Let y be an ordinal response with q categories, then the probability that y takes category k can be formed by the following cumulative logit odds (Agresti, 2010):

$$logit[P(y_{ij} \leq k|\boldsymbol{x})] = \mu_k + \boldsymbol{\beta'x}, \qquad k = 1, ..., q-1,$$

Here μ_k is the kth cut-off points of the response categories, i.e. the data is divided into kth cuts/ordinal categories. $\boldsymbol{x}$ represents a set of predictor variables which can be quantitative or categorical and $\boldsymbol{\beta}$ models the effect $\boldsymbol{x}$ has on the ordinal response. The proportional odds model then assumes the parameter vector $\boldsymbol{\beta'}$ is independent of category k and the parameter μ_k is dependent on category k. Consequently, the effects $\boldsymbol{\beta}$ yields on the ordinal response y_{ij} are the same across a set of predictors $\boldsymbol{x}$. In terms of interpretation, the proportional odds model describes the probability of being in one category (or lower) against the probability of being in higher categories.

Often, the proportional odds model is instead expressed as:

$$logit[P(y_{ij} \leq k|\boldsymbol{x})] = \mu_k - \boldsymbol{\beta'x}$$

The negative sign preceding the effect β for predictor x allows the natural interpretation of the effect β regarding whether it is positive or negative. More specifically, if $\beta_j > 0$ then higher values of x leads to higher values of Y and hence each cumulative logit decreases (Agresti, 2010). The general model with multiple explanatory variables satisfies:

$$logit[P(y_{ij} \leq k|\boldsymbol{x_1})] - logit[P(y_{ij} \leq k|\boldsymbol{x_2})] = \beta'(\boldsymbol{x_1} - \boldsymbol{x_2})$$

The odds of having response $y_{ij} \leq k$ for $\boldsymbol{x_1}$ is $exp(\beta'(\boldsymbol{x_1} - \boldsymbol{x_2}))$ times the odds of $\boldsymbol{x_2}$. In other words, the logit odds ratio is proportional to the distance between $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, hence the name *proportional odds model*. Define α_i and β_j as the individual row and column effects respectively, and the interaction effects of the individual rows and columns γ_{ij}. In this manner, the saturated model (Matechou et al., 2016) can be arranged as follows,

$$logit[P(y_{ij} \leq k|\boldsymbol{x})] = \mu_k - \alpha_i - \beta_j - \gamma_{ij} \tag{1}$$

for $k = 2, \ldots, q$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ with $\mu_1 < \mu_2 < \cdots < \mu_{q-1}$.

This model however is over-parameterized which means it requires a lot of parameters to describe the individual effects thus interpretation of the data becomes difficult. Hard clustering via finite mixtures is then introduced for reason of dimensionality reduction.

3.3.1 The proportional Odds Model with Clustering

The clustering formulae including the interaction term for the proportional odds model are then summarized as below:

- Row clustering,
$$logit[P(y_{ij} \leq k)] = \mu_k - \alpha_r - \beta_j - \gamma_{rj}$$
- Column clustering,
$$logit[P(y_{ij} \leq k)] = \mu_k - \alpha_i - \beta_c - \gamma_{ic}$$
- Biclustering,
$$logit[P(y_{ij} \leq k)] = \mu_k - \alpha_r - \beta_c - \gamma_{rc}$$

where $k = 1, \ldots, q$, $r = 1, \ldots, R$, $c = 1, \ldots, C$ and $i = 1, \ldots, n$

Here, α_r represents the row cluster effect and β_c represents the column cluster effects with $\sum_{r=1}^{R} \alpha_r = \sum_{c=1}^{C} \beta_c = 0$. Choosing $R \ll n$ and $C \ll m$ ensures that the number of independent/free parameters in this model is lower than the number of parameters in the proportional odds model in (1). The inclusion of the interaction term allows for different slopes and possible crossings where $\sum_r \gamma_{rj} = 0$, $\sum_c \gamma_{ic} = 0$, and $\sum_c \gamma_{rc} = 0 \quad \forall \; r, c$. The additive version of these models omits the interaction term. Note the difference between the biclustering models and the row clustering models: for biclustering, the models involve column cluster effects instead of individual column effects. Similarly the biclustering model involves row cluster effects, instead of the individual row effects used in the column clustering model.

The additive version of the row clustering model (excluding the interaction term) and its simpler version can be formed as below, respectively:

$$logit[P(y_{ij} \leq k)] = \mu_k - \alpha_r - \beta_j, \qquad logit[P(y_{ij} \leq k)] = \mu_k - \alpha_r$$

The column clustering and biclustering models can then be obtained in a similar manner. The probabilities that observation $y_{ij} = k$ (Matechou et al., 2016) in the row clustering case are given as follow:

$$\theta_{r_ijk} = \begin{cases} \frac{exp(\mu_k - \alpha_r - \beta_j - \gamma_{rj})}{1 + exp(\mu_k - \alpha_r - \beta_j - \gamma_{rj})}, & k = 1 \\ \frac{exp(\mu_k - \alpha_r - \beta_j - \gamma_{rj})}{1 + exp(\mu_k - \alpha_r - \beta_j - \gamma_{rj})} - \frac{exp(\mu_{k-1} - \alpha_r - \beta_j - \gamma_{rj})}{1 + exp(\mu_{k-1} - \alpha_r - \beta_j - \gamma_{rj})}, & 1 < k < q \\ 1 - \sum_{k=1}^{q-1} \theta_{r_ijk}, & k = q \end{cases}$$

where $k = 1, \ldots, q$, $r = 1, \ldots, R$, $c = 1, \ldots, C$ and $i = 1, \ldots, n$

In doing so, column clustering probabilities θ_{ic_jk} and biclustering probabilities $\theta_{r_ic_jk}$ are similar to that of row clustering for varying k, using different notations applicable in each clustering case.

Table 3.4 details the number of free parameters **npar** excluding π parameters for the proportional odds interaction model with clustering (Matechou et al., 2016). Thus, in order to obtain the number of free parameters that includes the π parameters, i.e. **npar**, have `v = npar + (`η_{g}` - 1)` where g can be replaced by π_r in row clustering, κ_c in column clustering, and π_r and κ_c for biclustering.

R	C	Model	npar
r	m	Row clustering	(q-1) + 2R + p - 4
n	c	Column clustering	(q-1) + 2C + n - 4
r	c	Biclustering	(q-1) + 2R + 2C - 4

Table 3.4: No.of.free.parameters for the interaction proportional odds model with clustering

3.4 Stereotype model

The general adjacent-categories logit model is:

$$logit[P(y_{ij} \leq k)] = \mu_k + \boldsymbol{\beta_j'}\boldsymbol{x}, \qquad k = 1, \ldots, q-1,$$

while the adjacent-categies logit model with proportional odds structure has the form

$$logit[P(y_{ij} \leq k)] = \mu_k + \boldsymbol{\beta'}\boldsymbol{x}, \qquad k = 1, \ldots, q-1$$

The stereotype model by definition is a multiplicative paired-category logit model that is nested between the above models (Agresti, 2010). The model is then:

$$log\frac{P(y_{ij} = k|x)}{P(y_{ij} = 1|x)} = \mu_k + \phi_k\boldsymbol{\beta'}\boldsymbol{x}$$

In terms of the category response probabilities, the stereotype model can be rewritten as:

$$P(y_{ij} = k|x) = \theta_{ijk} = \frac{exp(\mu_k + \phi_k\boldsymbol{\beta'}\boldsymbol{x})}{\sum_{l=1}^{q} exp(\mu_l + \phi_l\boldsymbol{\beta'}\boldsymbol{x})}, \qquad k = 1, 2, \ldots, q$$

with constraint $\mu_1 = \phi_1 = 0$. For logit k, the explanatory variable x_l has coefficients $\phi_k\beta_l$. For that reason, the odds ratio for a category k comparing with the baseline when there is a unit increase in x_l is $exp(\phi_k\beta_l)$. The $\{\phi_k\}$ parameters in the stereotype model can be regarded as scores for the outcome categories, allowing the model to treat the outcome variable as ordinal. These parameters are constrained for identifiability reasons, e.g. $\phi_1 = 0$ and $\phi_q = 1$. Another advantage of the stereotype model is that it can also be used to estimate the distance between two adjacent categories, e.g. k and $k+1$, based on how close their scores are, i.e. ϕ_k and ϕ_{k+1} (Costilla, 2017).

Since the stereotype model is equivalent to the adjacent-categories logit with proportional odds structure, it can be reformulated in terms of adjacent-categories logit as follows,

$$log\frac{P[y_{ij} = k|\boldsymbol{x}]}{P[y_{ij} = k+1|\boldsymbol{x}]} = (\mu_k - \mu_{k+1}) + (\phi_k - \phi_{k+1})\boldsymbol{\beta'}\boldsymbol{x} \tag{2}$$

$$= \tau_k + \nu_k\boldsymbol{\beta'}\boldsymbol{x}, \qquad k = 2, \ldots, q \tag{3}$$

The relation between $\{\phi_k\}$ and $\{\nu_k\}$ is defined by:

$$\nu_k = \phi_k - \phi_{k+1}, \qquad k = 1, \ldots, q,$$

and $\phi_k = \sum_{t=1}^{k-1} \nu_t, \quad k = 1, ..., q$. Therefore, the adjacent-categories logit model is a special case of the ordered stereotype model when $\boldsymbol{\nu_k = 1}$, i.e. when the $\{\phi_k\}$ scores are fixed and equally spaced (Fernández et al., 2016).

3.4.1 The Ordered Stereotype Model with clustering

With the increasing monotone constraint of ϕ_k

$$0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1,$$

the stereotype model can also be called as the ordered stereotype model. This implies that for a unit increase in the predictor variable x_l, the odds ratio $exp(\phi_k \beta_l)$ of category k against the baseline category increases when category k is further from the baseline. Using the same notations and sum-to-zero constraints for the row cluster effect α_r and the column effect β_c for the proportional odds model, the following formulae for the ordered stereotype model including clustering are as follows:

- Row Clustering

$$log \frac{P(y_{ij} = k|i \in r)}{P(y_{ij} = 1|i \in r)} = \mu_k - \phi_k(\alpha_r + \beta_j + \gamma_{rj})$$

- Column Clustering

$$log \frac{P(y_{ij} = k|j \in c)}{P(y_{ij} = 1|j \in c)} = \mu_k - \phi_k(\alpha_i + \beta_c + \gamma_{ic})$$

- Bilustering

$$log \frac{P(y_{ij} = k|i \in r, j \in c)}{P(y_{ij} = 1|i \in r, j \in c)} = \mu_k - \phi_k(\alpha_r + \beta_c + \gamma_{rc})$$

where $k = 1, \ldots, q$, $r = 1, \ldots, R$, $c = 1, \ldots, C$ $i = 1, \ldots, n$ and $j = 1, \ldots, m$

Choosing $R \ll n$ ($C \ll m$) makes certain that the number of independent parameters in these models is less than that of the saturated model. The probabilities that observation $y_{ij} = k$ given that row i belongs to row cluster r, and/or column j belongs to column c are expressed as:

- Row Clustering

$$\theta_{r_ijk} = P(y_{ij} = k|i \in r) = \frac{exp(\mu_k - \phi_k(\alpha_r + \beta_j + \gamma_{rj}))}{\sum_{l=1}^{q} exp(\mu_l - \phi_l(\alpha_r + \beta_j + \gamma_{rj}))}$$

- Column Clustering

$$\theta_{ic_jk} = P(y_{ij} = k|j \in c) = \frac{exp(\mu_k - \phi_k(\alpha_i + \beta_c + \gamma_{ic}))}{\sum_{l=1}^{q} exp(\mu_l - \phi_l(\alpha_i + \beta_c + \gamma_{ic}))}$$

- Biclustering

$$\theta_{r_ic_jk} = P(y_{ij} = k|i \in r, j \in c) = \frac{exp(\mu_k - \phi_k(\alpha_r + \beta_c + \gamma_{rc}))}{\sum_{l=1}^{q} exp(\mu_l + \phi_l(\alpha_r + \beta_c + \gamma_{rc}))},$$

where $k = 1, \ldots, q$, $r = 1, \ldots, R$, $c = 1, \ldots, C$ $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

The number of free parameters `npar` for the ordered stereotype interaction model with clustering (Fernández, 2015) is:

R	C	Model	npar
r	m	Row clustering	2q + Rm + (R-1) - 5
n	c	Column clustering	2q + Cn + (C-1) - 5
r	c	Biclustering	2q + RC + (R-1) + (C-1) - 5

Table 3.5: No.of.free parameters for the ordered stereotype interaction model with clustering

3.5 Binary model

In the binary case, we can view y_{ij} as a realization of a random variable $\mathbf{Y}$ that can take the values one and zero with probabilities ρ_{ij} and $1 - \rho_{ij}$, respectively. The probability distribution of $\mathbf{Y}$ is from a Bernoulli distribution with parameter ρ_{ij}, and can be written as:

$$P(Y = y_{ij}) = \rho_i^{y_{ij}}(1 - \rho_{ij})^{1-y_{ij}},$$

for $y_{ij} = 0, 1$. Note that for $y_{ij} = 1$ obtain ρ_{ij} and for $y_{ij} = 0$ obtain $1 - \rho_{ij}$. The odds ratio of favourable to unfavourable cases for ρ_{ij} and the logit or log-odds are defined as follows, respectively:

$$\text{odds}_{ij} = \frac{\rho_{ij}}{1 - \rho_{ij}}, \qquad o_{ij} = logit(\rho_{ij}) = log\frac{\rho_{ij}}{1 - \rho_{ij}}$$

Solving for ρ_{ij} returns $\rho_{ij} = logit^{-1}o_{ij} = \frac{e^{o_{ij}}}{1+e^{o_{ij}}}$. In terms of clustering, the rows and/or columns of the binary data matrix $\mathbf{Y}$ are modelled to be coming from finite mixtures groups with membership unknown, yielding a model-based hard clustering of the rows and/or columns of observations. The formulae for the binary model including clustering are then given as:

- Row clustering,
$$logit[y_{ij} = 1] = \mu + \alpha_r + \beta_j + \gamma_{rj},$$
- Column clustering,
$$logit[y_{ij} = 1] = \mu + \alpha_i + \beta_c + \gamma_{ic},$$
- Biclustering,
$$logit[y_{ij} = 1] = \mu + \alpha_r + \beta_c + \gamma_{rc}$$

where $r = 1, ..., R$, $c = 1, ..., C$ $i = 1, ..., n$ and $j = 1, ..., m$.

Using the same identifiability constraints imposed on α_r, α_i, β_c, β_j for the proportional odds and ordered stereotype models, i.e. $\sum_{r=1}^{R} \alpha_r = \sum_{c=1}^{C} \beta_c = 0$ and $\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{m} \beta_j = 0$, and for the interaction γ, $\sum_r \gamma_{rj} = \sum_c \gamma_{ic} = \sum_r \sum_c \gamma_{rc} = 0$ the probabilities that observation $y_{ij} = 1$ if $i \in r$ and/or $j \in c$ and $y_{ij} = 0$ otherwise are given as follows:

- Row Clustering

$$\theta_{ic_jk} = P(y_{ij} = 1 | i \in r) = \frac{exp(\mu + \alpha_r + \beta_j + \gamma_{rj})}{1 + exp(\mu + \alpha_r + \beta_j + \gamma_{rj})}$$

- Column Clustering

$$\theta_{r_ijk} = P(y_{ij} = 1|j \in c) = \frac{exp(\mu + \alpha_i + \beta_c\gamma_{ic} + \gamma_{rj})}{1 + exp(\mu + \alpha_i + \beta_c + \gamma_{rj})}$$

- Biclustering

$$\theta_{r_ic_jjk} = P(y_{ij} = 1|i \in r, j \in c) = \frac{exp(\mu + \alpha_i + \beta_c + \gamma_{rj})}{1 + exp(\mu + \alpha_r + \beta_c + \gamma_{rj})}$$

where $r = 1, ..., R$, $c = 1, ..., C$ $i = 1, ..., n$ and $j = 1, ..., m$.

The number of independent parameters `npar` for the binary interaction model including clustering is :(Pledger and Arnold, 2014)

R	C	Model	npar
r	m	Row clustering	Rm + R - 2
n	c	Column clustering	Cn + C - 2
r	c	Biclustering	RC + R + C - 3

Table 3.6: Number of free parameters for the interaction binary model with clustering

4 Likelihoods and Estimation of Parameters

The parameters for the finite mixture models with clustering may be estimated by the Expectation-Maximisation (EM) algorithm, with the incomplete or missing data being the unknown cluster membership of each row and/or column. Essentially, this procedure treats the unknown cluster membership as a missing data problem and then estimates the parameters using an iterative two-fold approach (Costilla, 2017). Since the likelihood surface is multimodal, the EM algorithm is started from a number of different points and the iteration with the highest likelihood value is obtained (Matechou et al., 2016).

EM algorithm procedure (Dempster et al., 1977; McLachlan and Krishnan, 2007):

1. Set starting values for the parameters.
2. **E-step**: Use the observed (but incomplete) data, the current parameter estimates and their expected values to estimate the missing data.
3. **M-step**: Numerically maximise the complete-data log-likelihood. Update the parameter estimates.
4. Stop the algorithm when it seems that the convergence has reached, that is, the change in the parameters is not significant.

The application of EM estimation on the clustering of rows and/or columns for the ordinal models are discussed in details in the next section (Costilla, 2017).

4.1 Row clustering

The overall (incomplete) likelihood for the row clustering model sums over all possible allocations of rows to row group:

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

The log likelihood does not have a simple form - it involves the log of a sum, which is inconvenient:

$$logL(\Omega, \pi|\mathbf{Y}) = \sum_{i=1}^{n} log \left\{ \sum_{r=1}^{R} \left[\pi_r \prod_{j=1}^{m} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij=k})} \right] \right\}$$

One possibility is to use the usual optimisation routine but this can be slow for the estimates to converge and finding a suitable set of starting points can be a problem. Alternatively, EM algorithm could be used instead as this is more stable in its approach to convergence.

As the actual membership of the rows among the R clusters and the columns among the, let z_{ir} be a random variable assigning data point i to its group. z_{ir} is a latent indicator variable taking value 1 if $i \in r$ otherwise value 0. Set $\mathbf{Z}$ as the missing data matrix and have the a priori row cluster proportions denoted as $\pi_1, ..., \pi_r$, with $\sum_{r=1}^{R} \pi_r = 1$. Given the observed but incomplete data matrix $\mathbf{Y}$ and the current parameter estimates Ω, the complete data log-likelihood is:

$$logL_C(\Omega, \boldsymbol{Y}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{r=1}^{R} \sum_{k=1}^{q} z_{ir} I(y_{ij} = k) log(\theta_{rjk}) + \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} log(\pi_r)$$

where Ω is obtained accordingly to each chosen model: $\Omega = (\mu, \alpha)$ for the proportional odds model, $\Omega = (\mu, \phi, \alpha)$ for the ordered stereotype model.

A priori z_{ir} follows a multinomial distribution:

$$z_i = (z_{i1}, \ldots, z_{iR})^T \sim Multinomial(1; \pi_1, \ldots, \pi_R), \qquad \sum_{r=1}^{R} z_r = 1.$$

Then a posteriori x_{jc} also follows a multinomial distribution:

$$z_i = (x_{i1}, \ldots, z_{iR})^T \sim Multinomial(1; \hat{z}_1, \ldots, \hat{z}_R)$$

In **E-step**, estimate the expected value of z_{ir}, $E[z_{ir}] = \hat{z}_{ir}$. This is also known as the posterior probabilities that row i is a member of row group r. Update $\mathbf{Z}$.

$$\hat{z}_{ir} = \frac{\pi_r \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij=k})}}{\sum_{g=1}^{C} \kappa_g \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{jgk}^{I(y_{ij=k})}}$$

In **M-step**, numerically maximise the complete log-likelihood. Find $\hat{\pi}_r$ and $\hat{\Omega}$ using the constraint $\sum_{r=1}^{R} \pi_r = 1$,

$$\hat{\pi}_r = \sum_{i=1}^{n} \frac{\hat{z}_{ir}}{n}, \qquad \hat{\Omega} = \underset{\Omega}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{r=1}^{R} \hat{z}_{ir} I(y_{ij} = k) log(\hat{\theta}_{rjk}) \right]$$

This two-step iteration repeats until parameter estimates have converged. The EM algorithm formulae for fitting the row clustering binary model are the similar to that of the ordinal model, replacing $f(\theta_{ij}, y_{ij}) = \theta_{rjk}^{I(y_{ij}=k)}$ with $f(\theta_{ij}, y_{ij}) = \theta_{rj}^{y_{ij}}(1-\theta_{rj})^{1-y_{ij}}$ and excluding the product and sum over q terms, with the current parameter estimate $\Omega = (\mu, \alpha)$. Additionally, the model formulation for column clustering is similar, with clustering of columns but not rows. The row clustering of the binary model, the column clustering of the ordinal and the binary models can be found in the appendices.

4.2 Biclustering

Introduce a $\boldsymbol{X \times 1}$ missing matrix $\mathbf{X}$ with $\boldsymbol{x_{jc}}$ as the latent column cluster membership for each column where $x_{jc} = 1$ of $j \in c$ and $x_{jc} = 0$ otherwise. The incomplete-data likelihood sums over all partitions of columns in C clusters is:

$$L(\Omega, \pi, \kappa|\mathbf{Y}) = \sum_{c_1=1}^{C} \cdots \sum_{c_m=1}^{C} \kappa_{c_1} \ldots \kappa_{c_m} \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \ldots \pi_{r_n} \prod_{i=1}^{n} \prod_{j=1}^{m} \prod_{k=1}^{q} \theta_{r_i c_j k}^{I(y_{ij}=k)}$$

The incomplete-data likelihood sums over all partitions of rows in R clusters is:

$$L(\Omega, \pi, \kappa|\mathbf{Y}) = \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \ldots \pi_{r_n} \sum_{c_1=1}^{C} \cdots \sum_{c_m=1}^{C} \kappa_{c_1} \ldots \kappa_{c_m} \prod_{i=1}^{n} \prod_{j=1}^{m} \prod_{k=1}^{q} \theta_{r_i c_j k}^{I(y_{ij}=k)}$$

Assuming row-based conditional independence and independence over the columns, and column-based conditional independence and independence over the rows, the corresponding incomplete-data log-likelihood are given as:

$$logL(\Omega, \pi, \kappa|\mathbf{Y}) = log\left\{ \sum_{c_1=1}^{C} \cdots \sum_{c_m=1}^{C} \kappa_{c_1} ... \kappa_{c_m} \prod_{i=1}^{n} \left[\sum_{r=1}^{R} \pi_r \prod_{j=1}^{m} \prod_{k=1}^{q} \theta_{r_i c_j k}^{I(y_{ij}=k)} \right] \right\}$$

$$logL(\Omega, \pi, \kappa|\mathbf{Y}) = log\left\{ \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \ldots \pi_{r_n} \prod_{j=1}^{m} \left[\sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{r_i c_j k}^{I(y_{ij}=k)} \right] \right\}$$

Consequently, the complete data log-likelihood can be simplified from using either the above incomple-data log-likelihood:

$$\ell_C(\Omega, Y, Z, X) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{r=1}^{R} \sum_{k=1}^{q} z_{ir} x_{jc} I(y_{ij} = k) log(\theta_{rck}) + \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ig} log(\pi_r) + \sum_{j=1}^{m} \sum_{c=1}^{C} x_{jc} log(\kappa_c)$$

In **E-step**, given $\mathbf{Y}$, the missing matrices $\mathbf{Z}$ and $\mathbf{X}$, and some chosen values for starting points Ω, the expected values of z_{ir} and x_{jc} are as follows,

$$\hat{z}_{ir} = \frac{\pi_r \prod_{j=1}^{m} \{\sum_{c=1}^{C} \kappa_c \prod_{k=1}^{q} \theta_{rck}^{I(y_{ij}=k)}\}}{\sum_{b=1}^{C} \kappa_c \prod_{i=1}^{n} \{\sum_{a=1}^{R} \pi_a \prod_{k=1}^{q} \theta_{abk}^{I(y_{ij}=k)}\}},$$

$$\hat{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n} \{\sum_{r=1}^{R} \pi_r \prod_{k=1}^{q} \theta_{rck}^{I(y_{ij}=k)}\}}{\sum_{b=1}^{C} \kappa_c \prod_{i=1}^{n} \{\sum_{a=1}^{R} \pi_a \prod_{i=1}^{n} \theta_{abk}^{I(y_{ij}=k)}\}}$$

Update **Z** and **X**.

In **M-step**, numerically maximising the complete-data log-likelihood to obtain the maximum likelihood estimates $\hat{\Omega}$, $\hat{\pi}_r$ and $\hat{\kappa}_c$,

$$\hat{\Omega} = \underset{\Omega}{\text{argmax}} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{r=1}^{R} \sum_{c=1}^{C} \hat{z}_{ir} \hat{x}_{jc} I(y_{ij} = k) log(\hat{\theta}_{rck}) \right], \qquad \hat{\pi}_r = \sum_{i=1}^{n} \frac{\hat{z}_{ir}}{n}, \qquad \hat{\kappa}_c = \sum_{j=1}^{m} \frac{\hat{x}_{jc}}{m}$$

The EM algorithm formulae for fitting the biclustering binary model again are the similar to that of the ordinal model, replacing $f(\theta_{ij}, y_{ij}) = \theta_{rjk}^{I(y_{ij}=k)}$ with $f(\theta_{ij}, y_{ij}) = \theta_{rj}^{y_{ij}}(1 - \theta_{rj})^{1-y_{ij}}$ and excluding the product and sum over q terms.

4.3 Conversion of the Score Parameters

Note that the starting values for ϕ do not correspond directly to ϕ, because ϕ is restricted to being increasing and between 0 and 1. Such a constraint is complex to impose during optimisation and so instead the starting values are treated as $u_k = logit(\phi_k)$ of a vector u which can be between $-\infty$ and $+\infty$ (Fernández, 2015).

$$-\infty \leq u_2 \leq u_3 \leq \cdots \leq u_{q-1} \leq +\infty$$

This makes the optimisation process more convenient and straight-forward since the new parameter vector u is more flexible. Once the MLEs of u_k is found they could be converted back to ϕ using the inverse of the logit function, $expit(x) = (1 + e^{-x})^{-1}$. More specifically (Fernández, 2015),

$$\phi_k = \begin{cases} 0, & k = 1 \\ expit(u_2), & k - 2 \\ expit(logit(\phi_2) + \sum_{l=3}^{k} e^{u_l}), & k = 3, \ldots, q - 1 \\ 1, & k = q \end{cases}$$

5 Model Selection

Model selection is a critical stage in cluster analysis necessary for statistical inference. Information criteria (CRI) method was proposed as an alternative to the Likelihood Ratio Tests (LRTs) method. By definition, CRI measures the loss of information as a balanced penalty taking into account how well the model fits the data (based on lack of fit) and the complexity of the model (based on lack of parsimony) (Fernández, 2015). The general formula is:

$$CRI = -2\ell + P$$

where ℓ is the maximised log-likelihood and P corresponds to the penalty term. A lower value of CRI indicates a better model where -2ℓ decreases and P increases when the number of parameters in the model increases, improving the fit. Therefore -2ℓ rewards goodness-of-fit whereas P discourages over-fitting because increasing the number of parameters in the model almost always improves the goodness of the fit. Thus an information criteria selects the best model in a set of candidate models. However, it does so even when all these models fit poorly without giving any warning of that, i.e. CRI doesn't measure the absolute quality of a model, but only the quality relative to other models. In this section, an overview of information criteria used for model selection are highlighted.

5.1 Akaike's Information Criterion (AIC)

AIC formula consists of a penalty term, $P = 2K$ where K is the number of free parameters,

$$AIC = -2\ell + 2K.$$

It is based on the assumption that there is an unknown true model and the aim is to select a candidate model that best approximates this true model. For a redundant set of few of different models with the same number of mixture components, adjustment can be made to select the model with higher likelihood among the redundant models. The use of AIC in choosing the correct number of free parameters has several disadvantages include failure of regularity conditions and over-estimation of the number of mixture components (mainly for multivariate normal distribution).

5.2 Modified AIC criterion for small sample size (AIC$_c$)

AIC's estimate is only valid asymptotically. In particular, when the number of parameters of estimate is large and the sample size is small, some correction for the criteria is thus required. The correction of AIC for a small number of data points is known as AIC$_c$,

$$AIC + \frac{2K(K+1)}{n-K-1}, \quad \text{where } \ P = \frac{2K(K+1)}{n-K-1}.$$

5.3 Bayesian information criterion (BIC)

BIC is based on the assumption that not only does an exactly true model exist, but this model is also included within the set of candidate models. In addition to this, BIC also assumes that the true model should only have less than five free parameters. BIC performs rather poorly when the true model has a complex structure and overly selects models that are too simple for real life applications. The below formula depicts BIC with n sample size and K free parameters,

$$BIC = -2\ell + Klog(n).$$

5.4 Integrated classification likelihood criterion (ICL)

ICL is a complete-likelihood-based information criterion and an approximation to BIC, which was proposed for clustering. The general formula of ICL is shown below:

$$ICL = -2\ell_c + Klog(n)$$

where ℓ_c is the log of the complete-data likelihood, K is the number of free parameters and n is the sample size. In terms of mixture modelling, ICL enhances the ability of the mixture model by providing clear patterns of clustering structure of the data. ICL can then be reformulated as:

$$ICL = -2\ell_c + Klog(n) + 2EN(S)$$

where $EN(S)$ is the entropy function and S is the number of clusters. $EN(S)$ measures the overlap of the mixture components and is defined by $EN(S) = \ell - \ell_c$. If the entropy is close to zero, the cluster structure will shows that the clusters are well-separated. On the contrary, clusters overlap if the entropy is large.

6 Data Application

The R package used for clustering the Ewave ordinal and binary data is the **clustord** package (McMillan, Louise and Fernández, Daniel and Matechou, Eleni, 2020). In short, the **clustord** consists of three main functions called **row clustering**, **column clustering** and **biclustering** functions. Since the Ewave data has 235 columns, the simplest version of the additive row and column clustering model (excluding the row and column individual effects) was implemented for clustering the Ewave data in order to avoid long computational time. Diverse combinations of independent starting points were proposed for fitting the models of different number of clusters using the EM algorithm. Only those with the best maximised likelihood value are reported back.

There are three sets of the proposed starting values for both the row and the column clustering models. All specific models then use whichever starting parameters corresponding to their formulae. For instance, Ordered Stereotype Model Row Clustering model of 2 clusters can be formed as:

$$log\frac{P(y_{ij}=k|i \in r)}{P(y_{ij}=1|i \in r)} = \mu_k - \phi_k(\alpha_r), \qquad k=1,2,3,4, \qquad r=1,2$$

Hence this model has a total of 6 starting parameters, μ_2, μ_3, μ_4, u_2, u_3 and α_1, i.e. $npar = 6$. Note here ϕ_2 ϕ_3 were replaced by u_2 and u_3 so these parameters can vary between $-\infty$ and ∞. The starting points of the stereotype model is then obtained from each set and thus run the model of 2 clusters three times. We repeated the same process for the other clustering models for different number of clusters. The true number of cluster for a specific model will have the the highest likelihood value. If adding another cluster to the current number of clusters results in one of the clusters having no observations then the current number of clusters is indeed the true number of clusters that best represents the data.

Table 6.3 outlines the numerical values of the parameters from the proposed sets of starting points. The 1^{st} set has the parameters μ_k, u_k, α_r and β_c evenly spaced ordered. The parameters from the 2^{nd} set are random. Finally, the α_r row cluster effect and β_c column cluster effect from the 3^{rd} set are larger than that from the other two sets, implying that the models should potentially cluster the data better with the 3^{rd} set.

1^{st}**set**	2^{nd}**set**	3^{rd}**set**
$\mu_2 = 0.4$, $\mu_3 = 0.6$, $\mu_4 = 0.8$, $\mu = 0.15$, $u_2 = 1.8$, $u_3 = 2$ $\alpha_1 = \beta_1 = 0.4, \alpha_2 = \beta_2 = 0.6$, $\alpha_3 = \beta_3 = 0.8, \alpha_4 = \beta_4 = 1.0$, $\alpha_5 = \beta_5 = 1.2, \alpha_6 = \beta_6 = 1.4$	$\mu_2 = 0.25$, $\mu_3 = 0.1$, $\mu_4 = 1.1$, $\mu = 0.15$, $u_2 = 0.7$, $u_3 = -0.5$ $\alpha_1 = \beta_1 = -2, \alpha_2 = \beta_2 = 0.08$, $\alpha_3 = \beta_3 = 0.01, \alpha_4 = \beta_4 = -0.15$, $\alpha_5 = \beta_5 = 0.64, \alpha_6 = \beta_6 = -1.$	$\mu_2 = 0.15$, $\mu_3 = -0.15$, $\mu_4 = 0.15$, $\mu = 0.15$, $u_2 = 0.25$, $u_3 = 0.5$, $\alpha_1 = \beta_1 = 3.5, \alpha_2 = \beta_2 = -3.0$, $\alpha_3 = \beta_3 = 2.5, \alpha_4 = \beta_4 = 4.5$, $\alpha_5 = \beta_5 = -2.0, \alpha_6 = \beta_6 = -5.0$

Table 6.1: Combinations of starting points

The data results are presented using tables, elbow plots and spaced mosaic plot. By definition, the function of the elbow plots is to plot the number of clusters against the AIC scores to determine the true number of clusters in the data. More specifically, the number of clusters with the lowest AIC value is the true number of clusters.

6.1 Ordered Stereotype Model: Data Visualization

The ordered stereotype model produced an empty cluster when the number of row clusters reached six for row clustering, and when the number of column clusters reached seven for column clustering. For this reason, we settled with the true number of clusters being five for row clustering and the true number of clusters being six for column clustering. **Table 6.2** shows that the information criteria are at the lowest and the maximised likelihood value is at the highest using the second set of starting point in both row clustering and column clustering cases. It is more clear that the true number of clusters for row clustering is five and the true number of clusters for column clustering is seven when inspecting the elbow plot, with the AIC values for these number of row and column clusters are much smaller than the other AIC values. However, the difference between the AIC score of C = 7 and the AIC score of C = 6 is not that significant.

2^{nd} set : $\mu_2 = 0.25$, $\mu_3 = 0.1$, $\mu_4 = 1.1$, $u_2 = 0.7$, $u_3 = -0.5$, $\alpha_1 = \beta_1 = -2$, $\alpha_2 = \beta_2 = 0.08$, $\alpha_3 = \beta_3 = 0.01$, $\alpha_4 = \beta_4 = -0.15$, $\alpha_5 = \beta_5 = 0.64$, $\alpha_6 = \beta_6 = -1.2$.

Model	R	C	npar	best.lli	AIC	AIC$_c$	BIC	ICL
Row Clustering, $\mu_k - \phi_k * \alpha_r$	2	1	6	-16574.44	33162.88	33162.88	33217.41	33080.36
	3	1	7	-16272.37	32562.64	32562.65	32632.75	32457.29
	4	1	8	-16284.72	32591.48	32591.5	32677.17	32467.31
	5	1	9	-16218.56	32463.09	32463.12	32564.37	32322.01
Column Clustering, $\mu_k - \phi_k * \beta_c$	1	2	6	-15786.61	31587.22	31587.23	31641.76	31525.32
	1	3	7	-15506.17	31030.33	31030.35	31100.45	30967.01
	1	4	8	-15426.43	30874.86	30874.88	30960.55	30796.86
	1	5	9	-15399.99	30825.84	30825.87	30927.12	30760.15
	1	6	10	-15386.63	30801.08	30801.08	30917.9	30746.73
	1	7	11	-15384.39	30787.69	30787.73	30920.13	30735.55

Table 6.2: Ordered Stereotype Clustering Model

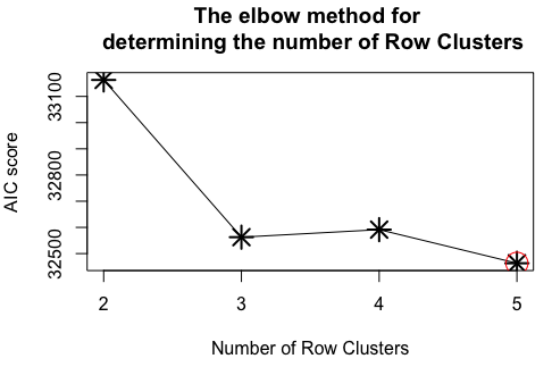

Figure 1: Elbow Plot for Row Clusters

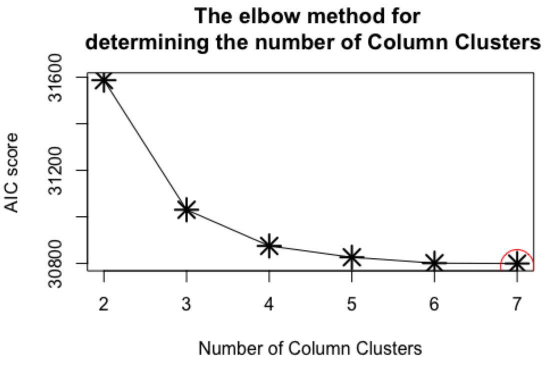

Figure 2: Elbow Plot for Column Clusters

6.2 Proportional Odds Model: Data Visualization

Table 6.3 shows that the information criteria are at the lowest and the maximised likelihood value is at the highest using the first set of starting point in both clustering cases. The proportional model also produces an empty cluster when the number of row clusters reached six for row clustering, and when the number of column clusters reached seven for column clustering. Based on this information and the numerical values of the maximised likelihood and information criteria from table 6.3, it can be definite that the true number of row clusters that best represent the data according to the proportional odds model is also five. However, since some of the information criteria for C = 6 such as AIC$_c$ and BIC are slightly smaller than these for C = 7, the true number of clusters for column clustering can be either 6 or 7. In the elbow plot below, it is very clear in the row clustering case that the true number of clusters in indeed five. Elbow plot however is more ambiguious when determining which number of column clusters between C = 6 and C = 7 is closer to the true number of clusters in the column clustering case.

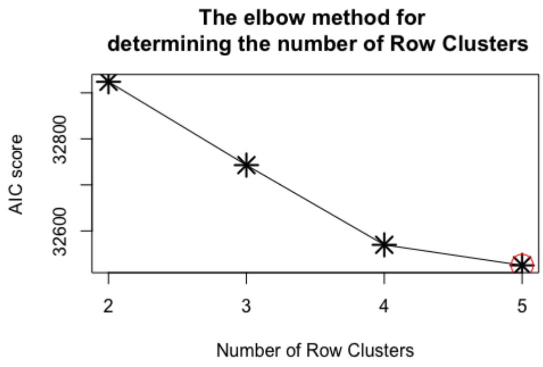

Figure 3: Elbow Plot for Row Clusters

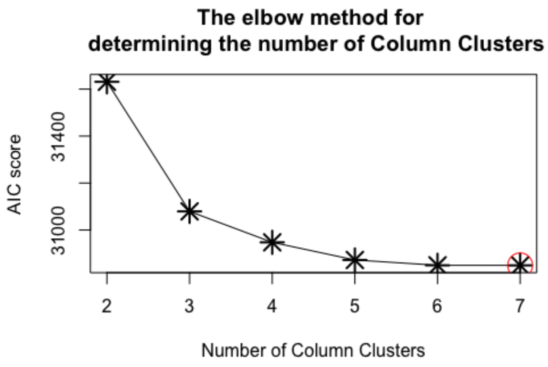

Figure 4: Elbow Plot for Column Clusters

$1^{st} set:\ \mu_2 = 0.4,\ \mu_3 = 0.6,\ \mu_4 = 0.8,\ u_2 = 1.8,\ u_3 = 2,\ \alpha_1 = \beta_1 = 0.4,\ \alpha_2 = \beta_2 = 0.6,\ \alpha_3 = \beta_3 = 0.8,\ \alpha_4 = \beta_4 = 1.0,\ \alpha_5 = \beta_5 = 1.2,\ \alpha_6 = \beta_6 = 1.4.$

Model	R	C	npar	best.lli	AIC	AIC$_c$	BIC	ICL
Row Clustering, $\mu_k - \alpha_r$	2	1	4	-16456.96	32923.92	32923.93	32962.88	32872.57
	3	1	5	-16364.49	32756.35	32756.36	32810.88	32681.55
	4	1	6	-16251.63	32569.83	32569.84	32639.94	32476.49
	7	1	9	-16218.56	32525.26	32525.28	32610.96	32408.78
Column Clustering, $\mu_k - \beta_c$	1	2	4	-15810.52	31631.03	31631.04	31669.98	31598.22
	1	3	5	-15532.67	31079.34	31079.35	31133.87	31035.39
	1	4	6	-15464.81	30947.62	30947.63	31017.73	30904.12
	1	5	7	-15425.22	30849.76	30872.68	30958.36	30844.76
	1	6	8	-15411.85	30849.76	30849.79	30951.04	30824.71
	1	7	9	-15409.82	30850.49	30850.49	30967.32	30823.25

Table 6.3: Proportional Odds Clustering Model

6.3 Binary Model : Data Visualization

For binary model, the information criteria are at the lowest and the maximised likelihood value is at the highest using the first set of starting point in both clustering cases. The binary model also produces an empty cluster when the number of row clusters reached six for row clustering, and when the number of column clusters reached seven for column clustering, similarly to that of the ordered stereotype model and the proportional odds model. It seems the information criteria from Table 6.4 suggest that the true number of column clusters is six, but the maximised likelihood value suggests otherwise though the difference of the likelihood between C = 6 and C = 7 is not significant. Thus the true number of column clusters is potentially 6. For row clustering, the true number of row clusters is also five, as seen in the elbow plot, with much lower values of information criteria and higher value of the maximised likelihood.

$1^{st} set$: $\mu_2 = 0.4$, $\mu_3 = 0.6$, $\mu_4 = 0.8$, $u_2 = 1.8$, $u_3 = 2$, $\alpha_1 = \beta_1 = 0.4$, $\alpha_2 = \beta_2 = 0.6$, $\alpha_3 = \beta_3 = 0.8$, $\alpha_4 = \beta_4 = 1.0$, $\alpha_5 = \beta_5 = 1.2$, $\alpha_6 = \beta_6 = 1.4$.

Model	R	C	npar	best.lli	AIC	AIC$_c$	BIC	ICL
Row Clustering, $\mu + \alpha_r$	2	1	2	-10644.02	21294.04	21294.04	21317.41	21264.79
	3	1	3	-10577.47	21164.94	21164.94	21203.89	21122.62
	4	1	4	-10551.08	21116.16	21116.16	21170.68	21048.87
	5	1	5	-10540.34	21098.69	21098.7	21168.8	21017.25
Column Clustering, $\mu - \beta_c$	1	2	2	-10059.1	20124.2	20124.2	20147.57	20107.39
	1	3	3	-9838.708	19687.42	19687.42	19726.37	19677.46
	1	4	4	-9755.979	19526.01	19526.01	19580.54	19515.1
	1	5	5	-9719.076	19456.15	19456.16	19526.26	19438.92
	1	6	6	-9713.986	19448.56	19448.58	19534.25	19453.92
	1	7	7	-9712.223	19451.13	19451.16	19552.41	19448.07

Table 6.4: Binary Clustering Model

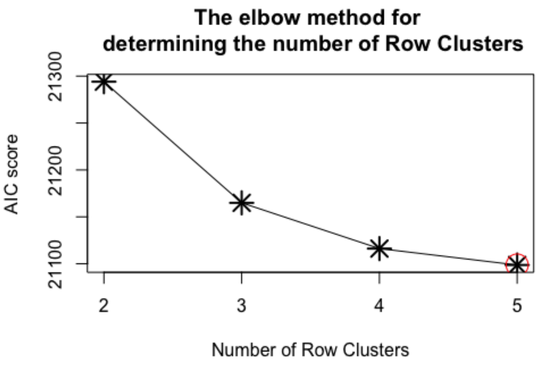

Figure 5: Elbow Plot for Row Clusters

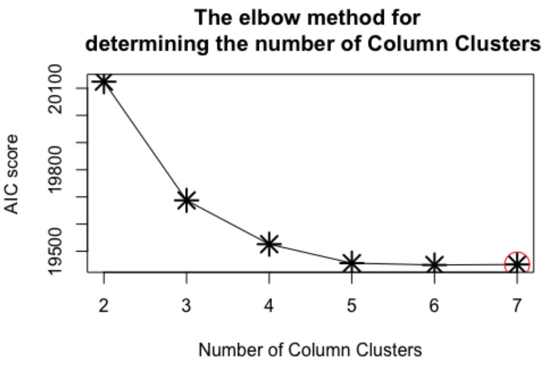

Figure 6: Elbow Plot for Column Clusters

6.4 Comparison between the models

Table 6.5 summarizes the candidate models along with their maximised likelihood values and information criteria values. The binary model of six column clusters using the first set of starting points appears to be the best model amongst the candidate models with much lower values of information criteria and significantly larger values of likelihood. This could be due to the binary model's small number of free parameters in comparison to that for the other two models. In addition, the ordered stereotype model (OSM) seems to be a better fit than the proportional odds model (POM) in both clustering cases with smaller information criteria values, though not significantly.

Model	Number of Clusters	best.lli	AIC	AIC$_c$	BIC	ICL
OSM	R = 5	-16218.56	32463.09	32463.12	32564.37	32322.01
	C = 7	-15384.39	30787.69	30787.73	30920.13	30735.55
POM	R = 5	-16218.56	32525.26	32525.28	32610.96	32408.78
	C = 7	-15409.82	30850.494	30850.49	30967.32	30823.25
Binary	R = 5	-10540.34	21098.69	21098.74	21168.8	21017.25
	C = 6	-9713.986	19448.56	19448.58	19534.25	19453.92

Table 6.5: Summary of the candidate models

Model	R1	R2	R3	R4	R5
OSM	1, 2, 4, 8, 9, 10, 11, 39, 41, 47, 48, 51, 53, 55, 64, 67, 72, 76	54, 60, 66	7, 13, 14, 15, 16, 20, 21, 23, 24, 25, 27, 30, 34, 35, 38, 42, 44, 59, 61, 65, 68, 69, 73	3, 5, 6, 12, 17, 18, 19, 22, 26, 28, 29, 32, 33, 36, 37, 40, 43, 45, 46, 49, 50, 52, 56, 57, 58, 62, 63, 71, 74, 75	31, 70
POM	3, 5, 6, 11, 12, 17, 18, 19, 22, 26, 28, 29, 32, 33, 36, 37, 40, 43, 45, 48, 49, 50, 52, 56, 57, 58, 62, 67, 71, 74, 75	7, 13, 14, 15, 16, 20, 21, 23, 24, 25, 27, 30, 34, 35, 38, 42, 44, 59, 61, 63, 65, 68, 69, 73	54, 60, 66	31, 70	1, 2, 4, 8, 9, 10, 39, 41, 47, 51, 53, 55, 64, 72, 76
Binary	3, 5, 6, 11, 14, 18, 19, 25, 26, 28, 29, 42, 43, 44, 47, 48, 50, 56, 57, 58, 59, 62, 63, 71	7, 12, 17, 21, 24, 30, 34, 35, 37, 49, 52, 61, 65, 68, 74, 75	16, 66	13, 15, 20, 23, 27, 38, 54, 73	1, 2, 4, 8, 9, 10, 22, 31, 32, 33, 36, 39, 40, 41, 45, 46, 51, 53, 55, 60, 64, 67, 69, 70, 72, 76

Table 6.6: Clustering structures for the Ordered Stereotype Model (OSM), Proportional Odds Model (POM) and Binary Model in the row clustering case

From **Table 6.5**, it has shown the column clustering model for the ordinal and the binary data overall has returned better results of the maximised likelihood values and the information criteria values than that for the row clustering model. In summary, the ordered stereotype model fits the data well using the second set of starting points whereas the proportional odds and the binary models fits the data better using the first set of starting points in terms of clustering.

Table **6.6** shows which row groups the rows are allocated into for the ordinal models and the binary model. R1, R2, R3, R4 and R5 are the row cluster groups where the integers in each cluster as seen in the table are the 76 observations or varieties of English. The clustering structures show the stereotype model and the proportional odds model both assign observations 31 and observation 70 into the same group. However, the label of the group of these observations is different for each model where the label for the stereotype model is R5 and the label for the proportional odds model is R4. The label of the row cluster consisting of observation 54, observation 60 and observation 66 for the stereotype model is also different to that for the proportional odds model (R3 for OSM, R2 for POM). This is known as label switching.

Label switching in finite mixture is a common problem for both Bayesian inference and EM algorithm methods. It is the non-identifiability of the labels in the mixture components or clusters where the labels of the cluster get mixed up when using different models to fit the data. Label switching is also implied when the allocation of individuals into clusters is different between different models. For example, the allocation of rows for the binary model is shown to be entirely different to that for the ordinal models despite using the same data. Label switching also occurs in the column clustering case.

6.5 Spaced Mosaic Plot

There are many visualization tools which depict reduction of dimensionality in matrices of ordinal data. In previous studies, spaced mosaic plot was introduced (Fernández et al., 2014) through graphing ordinal data for the ordered stereotype model. The examples of mosaic plots in this section were produced using an R function with permission called `spaced.mosaic.plot` (Fernández et al., 2014).

The Ewave data matrix without missing data has `12 rows and 235 columns` hence there are 12 rows × 235 columns = 2820 ordinal responses or cells in total. **Figure 7** shows the overall distribution of ordinal responses across all the cells, while ignoring rows and columns. Thus, the area of each block is equivalent to frequency. The ordinal category response 1 is most common by far, and ordinal category response 4 is the least. In other words, most of the 235 features are present in the majority of 76 varieties of English.

Figure 8 shows the clustering in the rows, allocating each row into one of the specified five clusters, basing on the distribution of ordinal responses across the columns of the data matrix. The height of each row group is proportional to the number of rows in the group. Within each row group we represent the frequencies of the four ordinal categories by the area of each block. Members of all row groups show very strong preference for category 1, and very weak preference for category 4 (all except for members from R4). In particular, R2 for category 1 is the biggest row cluster, consists of many cells representing the varieties of English (row) that have every grammatical features present. In contrast, only 14 cells with category 4 lie in R1. This means the minority of the varieties of English in R1 rarely has grammatical features that don't present.

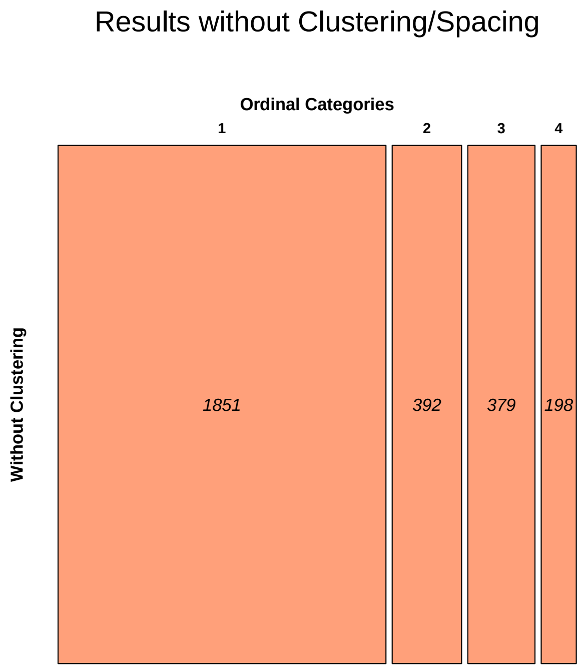

Figure 7: Mosaic plot without spacing or clustering

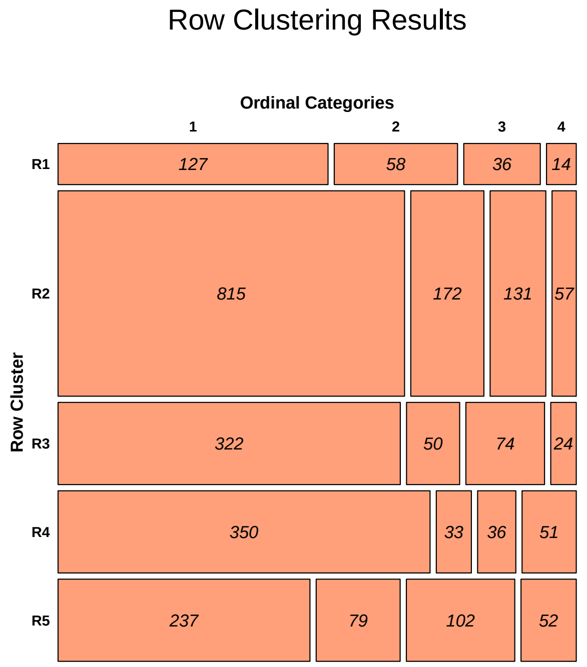

Figure 8: Mosaic Plot with Row Cluster for R=4

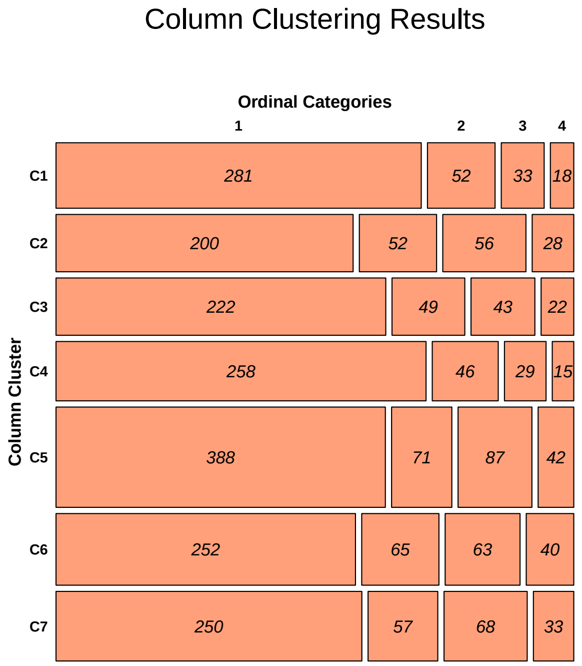

Figure 9: Mosaic Plot with Column Cluster for C=7

On the contrary, **Figure 9** shows the pooling of frequencies of all the ordinal responses for 235 features in the same column groups. The attribute of the column mosaic plot is very similar to that of the row mosaic plot, where the height of each column group is proportional to the number of columns in the group. For example, column groups C4 and C6 have the most cells while column group C2 has the least cells across all categories. Members or features of all column groups have very strong preference for category 1, where more than half of the cells are shown to be in favour of category 1. The rest of the cells are divided between categories 2, 3 and 4. The idea of column clustering is very useful to apply to the clustering of the Ewave data for we are interested in finding the similarities and differences between the features basing on the different categories. Overall, it is desirable to differentiate the grammatical features into groups in order to determine the characteristics of the features and answer the following questions. Which features are always present? Which features are common in the varieties of English? Is there any special feature that is very rare for some varieties but very common for others?

Figure 10 separates the row groups across the ordinal response categories showing the distances between the categories that the model has identified, i.e. distances between the adjacent score parameters, ϕ_k. The ϕ_k parameters identified are $\phi_1 = 0, \phi_2 = 0.4, \phi_3 = 0.6, \phi_4 = 1$. The distance between ϕ_1 and ϕ_2 (0.2) is relatively smaller than the distance between ϕ_1 and ϕ_2 (0.4), where the distance between ϕ_3 and ϕ_4 is the same as the distance between ϕ_1 and ϕ_2. The coloured lines placed between the row groups of different categories described the spaces between these categories. In that manner, the distances between the categories can be observed using these colour lines without having to refer to the values of ϕ_k. It seems that all the categories are differentiated from each other very well. The spaced mosaic plot for column clustering shows very similar results to the that for row clustering.

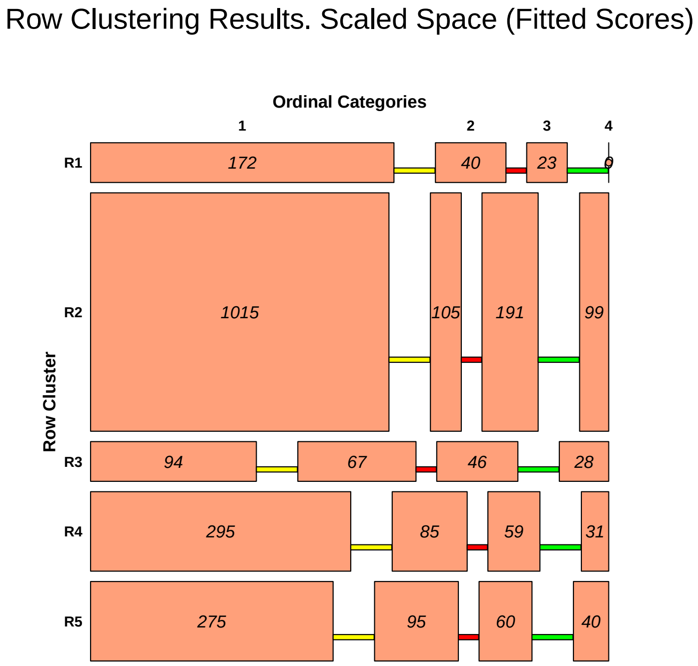

Figure 10: Spaced Mosaic Plot with clustering for R=5

7 Simulation Study

7.1 Design of the study

Goal:

We were interested in testing the reliability of the EM algorithm's performance in estimating the model parameters, i.e. check whether the estimated parameters are close to the true parameters.

Objectives:

- Perform the simulation study of the row clustering model on proportional odds model, ordered stereotype model and binary model for N iterations, using a different combination of starting points each iteration
- Obtain the simulated maximised likelihood estimates over N iterations and their standard error values, and finally compare these values to the true parameters.

The procedure for simulating the EM algorithm in the row clustering case is then outlined in the following steps:

Step 1. Generate some pseudo-data represented as a Y matrix (NRows×NCols), where the number of categories is $q = 4$. The pseudo-data is of known R clusters, where $R = 2$ with n_r cluster size. Thus $\mathbf{n_1 + n_2 = NRows}$. The elbow plot below shows that the EM algorithm has correctly estimated the true number of clusters in the data, which is $R = 2$.

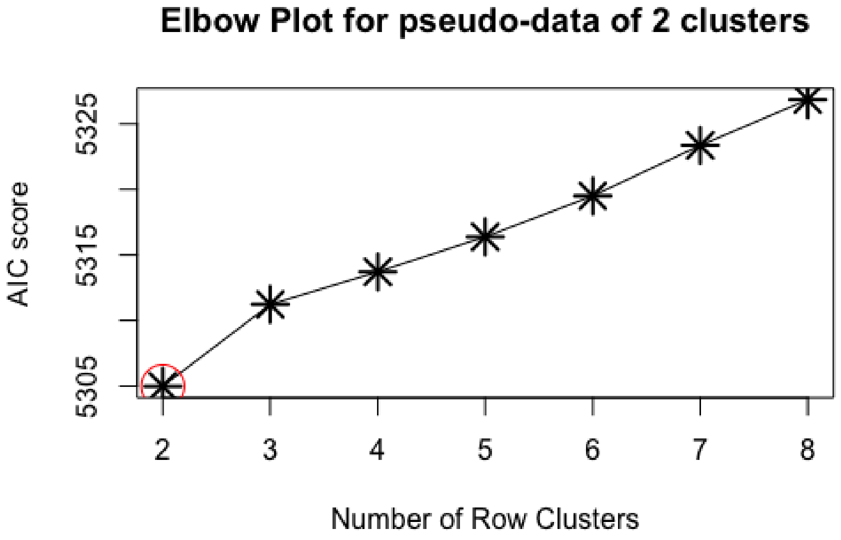

Figure 11: Elbow Plot for 2 Clusters

Step 2. Specify the simulation factors. Note that *npar* does not include the row membership proportion, π_r and that it should be selected after having already specified the model. As increasing the number of columns will add more information and effect for the response variable, the number of columns NCols will affect the outcome of the ordinal and binary response variable.

No.	Simulation Factor	Levels
1	Model	"OSM", "POM" and "Binary"
2	Formula	$\mu_k - \phi_k \alpha_r$, $\mu_k - \alpha_r$, $\mu - \alpha_r$
3	No.of.Columns, NCols	NCols $\in \{10, 20, 30\}$
4	Sample size of pseudo-data clusters, n_r where r=1,2	$n_r \in \{15, 50, 100\}$
6	No.of.free.parameters, npar	(q-1) + (q-2) + (R-1) (q-1) + (R-1) (R-1) + 1

Table 7.1: The simulation factors of the EM algorithm simulation study

Step 3. Choose the true parameter values for the specified models used to generate the two clusters in the pseudo-data. For ordered stereotype model, fix the following constraints, $0 = \phi_1 \leq \cdots \leq \phi_q = 1$ and $\mu_1 = \phi_1 = 0$. As for the proportional odds model, fix the constraints, $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{q-1}$. Finally, fix $\sum_{r=1}^{R} \alpha_r = 0$ for all models. The true parameters can be anywhere from $-\infty$ to ∞, taking the constraints into account.

Step 4. Run the simulation for N = 100 iterations in order to obtain the estimates and their standard errors, and compare these simulated estimates to the true parameter estimates. The standard errors is caculated using `cal.SE.rowcluster` function from the `clustord` package (McMillan, Louise and Fernández, Daniel and Matechou, Eleni, 2020).

We generated another data of four clusters to see if the EM algorithm is consistent in estimating the model parameters. The results however shows that the true number of clusters was estimated to be R = 2 instead of R = 4. This shows that the EM algorithm did not perform well for data of four clusters, despite the fact that the algorithm correctly estimated the true number of clusters for the pseudo-data of 2 clusters.

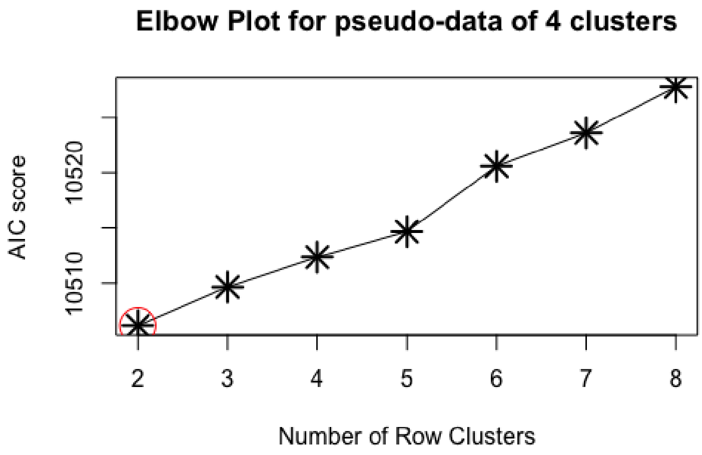

Figure 12: Elbow Plot for 4 Clusters

7.2 Results

			NCols = 30					
			NRows = 30		NRows = 100		NRows = 200	
			$n_1 = 15$, $n_2 = 15$		$n_1 = 50$, $n_2 = 50$		$n_1 = 100$, $n_2 = 100$	
Model	npar	True Estimate	Estimate	S.E	Estimate	S.E	Estimate	S.E
OSM	6	μ_2 = 0.6	0.589	0.380	0.580	0.286	0.552	0.199
		μ_3 = 0.8	0.693	0.498	0.729	0.336	0.733	0.231
		μ_4 = 1.2	0.879	0.452	1.088	0.268	1.147	0.192
		ϕ_2 = 0.4	0.487	NaN	0.439	NaN	0.333	NaN
		ϕ_3 = 0.6	0.588	NaN	0.571	NaN	0.573	NaN
		α_1 = 0.5	-0.045	0.548	-0.047	0.339	-0.070	0.241
POM	4	μ_2 = 0.6	-0.554	0.125	-0.555	NaN	-0.572	NaN
		μ_3 = 0.8	0.056	0.132	0.054	NaN	0.023	NaN
		μ_4 = 1.2	0.304	0.187	0.298	NaN	0.263	NaN
		α_1 = 0.5	-0.004	0.167	-0.011	NaN	-0.005	NaN
Binary	2	μ = 0.6	0.176	NaN	0.178	NaN	0.114	NaN
		α_1 = 0.5	-0.012	NaN	-0.008	NaN	0.024	NaN

Table 7.2: Simulated Estimates and their S.Es for NCols=30

			NCols = 10					
			NRows = 30		NRows = 100		NRows = 200	
			$\mathbf{n_1 = 15,\ n_2 = 15}$		$\mathbf{n_1 = 50,\ n_2 = 50}$		$\boldsymbol{n_1 = 100,\ n_2 = 100}$	
Model	npar	True Estimate	Estimate	S.E	Estimate	S.E	Estimate	S.E
OSM	6	$\boldsymbol{\mu_2}$ = 0.6	0.612	0.126	0.705	0.117	0.482	0.199
		$\boldsymbol{\mu_3}$ = 0.8	0.852	0.145	0.778	0.120	0.584	0.231
		$\boldsymbol{\mu_4}$ = 1.2	1.468	0.132	1.064	1.083	0.902	0.192
		$\boldsymbol{\phi_2}$ = 0.4	0.587	NaN	0.642	NaN	0.609	NaN
		$\boldsymbol{\phi_3}$ = 0.6	0.594	NaN	0.955	NaN	0.655	NaN
		$\boldsymbol{\alpha_1}$ = 0.5	0.026	0.339	-0.017	NaN	-0.044	NaN
POM	4	$\boldsymbol{\mu_2}$ = 0.6	-0.408	0.303	-0.504	0.158	-0.549	0.130
		$\boldsymbol{\mu_3}$ = 0.8	0.104	0.303	0.131	NaN	0.158	0.130
		$\boldsymbol{\mu_4}$ = 1.2	0.402	0.303	0.396	NaN	0.158	0.131
		$\boldsymbol{\alpha_1} = 0.5$	0.001	0.403	-0.012	0.178	-0.015	0.151
Binary	2	$\boldsymbol{\mu}$ = 0.6	0.108	NaN	0.184	NaN	0.126	NaN
		$\boldsymbol{\alpha_1}$ = 0.5	-0.040	NaN	-0.034	NaN	0.014	NaN
			NCols = 20					
			NRows = 30		NRows = 100		NRows = 200	
			$\mathbf{n_1 = 15,\ n_2 = 15}$		$\mathbf{n_1 = 50,\ n_2 = 50}$		$\boldsymbol{n_1 = 100,\ n_2 = 100}$	
Model	npar	True Estimate	Estimate	S.E	Estimate	S.E	Estimate	S.E
OSM	6	$\boldsymbol{\mu_2}$ = 0.6	2.502	0.306	0.581	0.126	0.600	0
		$\boldsymbol{\mu_3}$ = 0.8	2.932	0.296	0.699	0.145	0.852	NaN
		$\boldsymbol{\mu_4}$ = 1.2	3.293	0.337	1.067	0.132	1.262	NaN
		$\boldsymbol{\phi_2}$ = 0.4	0.759	NaN	0.558	NaN	0.481	NaN
		$\boldsymbol{\phi_3}$ = 0.6	0.904	NaN	0.570	NaN	0.742	NaN
		$\boldsymbol{\alpha_1}$ = 0.5	0.530	0.349	-0.039	0.1444	-0.036	NaN
POM	4	$\boldsymbol{\mu_2}$ = 0.6	-0.487	0.117	-0.672	0.140	-0.544	0.131
		$\boldsymbol{\mu_3}$ = 0.8	0.054	0.120	-0.050	0.141	0.055	0.131
		$\boldsymbol{\mu_4}$ = 1.2	0.352	NaN	0.222	0.141	0.298	0.131
		$\boldsymbol{\alpha_1} = 0.5$	-0.001	0.151	0.007	NaN	-0.011	0.140
Binary	2	$\boldsymbol{\mu}$ = 0.6	0.187	NaN	0.168	NaN	0.134	NaN
		$\boldsymbol{\alpha_1}$ = 0.5	-0.008	NaN	-0.05	NaN	0.052	NaN

Table 7.3: Simulated estimates $\{\mu_k,\ \mu,\ \phi_k,\ \alpha_r\}$ and their S.Es for NCols = 10, 20 and NRows = 30, 100, 200 for q = 4 categories. The row clustering model used is $\boldsymbol{\mu_k - \phi_k * \alpha_r}$, $\boldsymbol{\mu_k - \alpha_r}$ and $\boldsymbol{\mu - \alpha_r}$ for the ordered stereotype model, proportional odds model and the binary model, respectively. Some of the S.E values returned NaN, which implies a result that cannot be calculated, or is not a floating point number.

The simulated estimates in blue are the estimates that are close to the true parameter values. Table 7.2 results shows the EM algorithm's ability in estimating the parameters correctly for the ordered stereotype model, especially for the (NRows=30, NCols=10) case, the (NRows=100, NCols=10) case and the (NRows=30, NCols = 30) case. The alpha estimates however in this models however were not estimated correctly. According to this simulation study, the EM algorithm did not estimate the proportional odds model's and the binary model's parameters correctly. The row clustering structure for the binary model also produced an empty cluster out of two clusters for the pseudo-data, meaning the simulation R function does not work properly for the binary data. Moreover, binary model is the only model that returns all NaN values for the standard error calculations.

8 Discussion

We have presented the definitions and uses of several finite mixture likelihood based models for the linguistics ordinal and binary data in the concept of hard clustering. Hard clustering via finite mixtures was recommended in order to reduce the dimensionality of the saturated models and hence simplify the interpretation of the results. Additionally, the likelihood model-based approach has an advantage of providing the maximum likelihood estimates of the model parameters as well as model selection for statistical inference. The ordinal models we have described with clustering are the proportional odds and the ordered stereotype models. These models have the advantage of having fewer parameters to interpret in comparison to other ordinal models on the account of their parsimonious and proportional odds structure. When paired with clustering, the ordinal models then become even more parsimonious and flexible.

The procedure of the EM algorithm in fitting the models and estimating the model parameters in the clustering context was reported as an important stage of the cluster analysis. Different combinations of starting points were used to avoid convergence to a local maximum. Moreover, we set up a simulation study for testing the reliability of the EM algorithm method in correctly estimating the model parameters' true values. The design of the experiment used the artificially generated data with the known number of clusters and the aim was to confirm whether this estimation method would correctly estimate the maximised likelihood estimators for different row clustering models. In general, the results showed that the EM algorithm has correctly estimated all model true parameters except for the row cluster effect (α_r) parameters for the ordered stereotype model. The results however were proving that the EM algorithm did not correctly estimated the true parameters for the proportional odds and the binary models.

We have described the task of model selection in selecting the valid models with the correct number of clusters amongst a set of candidate models. The tables show the information criteria for the row clustering models are at the lowest when R = 5. For the column clustering models, the information criteria are at the lowest when C = 6 or C = 7. The column clustering models in general were performing better than the row clustering models, returning higher values of maximised likelihood function and lower values of information criteria. This could be due to the fact that the Ewave data has many columns and not so many rows. The best model according to the information criteria is then the binary model with column clustering of C = 6 column clusters as the true number of clusters that best represent the data.

Although it was concluded that the binary model is the best model, we used the ordered stereotype model to present data visualization methods with the aim is to describe the outputs from modelling the Ewave linguistics data. This is because binary data has less information than ordinal data, hence interpretation for ordinal data is expected to be more thoroughly and in depth. In particular, spaced mosaic plots focusing on the ordered stereotype models were used to describe the Ewave data in order to answer to the research questions.

Data visualization allows the display the distribution of the four categories for the response variable of the Ewave data, where category 4 (coded for absent) is the least distributed and category 1 (coded for present) is the most distributed. Category 2 (coded for common) and category 3 (coded for rare) are evenly distributed, approximately. From observing the mosaic plot for the column clustering model and row clustering model, it seems that cluster C5 is the largest out of all the column clusters and cluster R2 is the largest out of all the row clusters. Technically, the rows were clustered using the columns and the column were clustered using the rows. In terms of the Ewave data set, this implies the varieties of English in a row cluster share some similarities basing on the grammatical features of the English language, and the different English features share some similarities basing on the varieties of English.

Future research will develop an R function to graph the clustering structures of the ordered stereotype model, the proportional odds model and the binary model. This function should be able to graph the clustering results obtained directly from the clustering functions of the **clustord** package and to show which observations are in which groups for the sake of a more comprehensive interpretation. So far only the additive models without the column and row individual effects were attempted for reasons of computationally extensivity. In the future, we wish to attempt other versions of the clustering models including the interaction term plus the biclustering models for determining the true number of clusters in the Ewave data. More varying combinations of starting points should also be considered for running the EM algorithm for better results might be achieved. In related to this matter, another future research goal is to run the simulation study with more clustering models for longer N iterations in order to produce better results. Most importantly, we wish to investigate the problem behind the simulation function used for simulating binary pseudo-data, as well as the issue from the **calc.SE.rowcluster** function from the **clustord** package. Other general future direction could be to develop a simulation study with the purpose of evaluating and comparing the performance of different model selection criteria with ordinal and binary data for the finite mixture modelling method. In doing so we will be able to see which information criteria correctly determine the true number of clusters by calculating the percentage of total simulated experiments where each criteria accurately estimates the true number of clusters for different scenarios. In addition, it would be helpful to look into the estimation of the model parameters in a Bayesian inference framework and compare that to the EM algorithm method so decide the advantages and disadvantages of the two methods.

Appendices

A Row clustering for the binary models

The overall (incomplete) likelihood for the row clustering model sums over all possible allocations of rows to row group:

$$L(\Omega, \pi|\mathbf{Y}) = \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \left[\pi_r \prod_{j=1}^{m} \theta_{rj}^{y_{ij}} (1-\theta_{rj})^{1-y_{ij}} \right] \right\}$$

The log likelihood does not have a simple form - it involves the log of a sum, which is inconvenient:

$$logL(\Omega, \pi|\mathbf{Y}) = \sum_{i=1}^{n} log \left\{ \sum_{r=1}^{R} \left[\pi_r \prod_{j=1}^{m} y_{ij} log(\theta_{rj}) + (1-y_{ij}) log(1-\theta_{rj}) \right] \right\}$$

As the actual membership of the rows among the R clusters and the columns among the, let z_{ir} be a random variable assigning data point i to its group. z_{ir} is a latent indicator variable taking value 1 if $i \in r$ otherwise value 0. Set $\mathbf{Z}$ as the missing data matrix and have the a priori row cluster proportions denoted as $\pi_1, ..., \pi_r$, with $\sum_{r=1}^{R} \pi_r = 1$. Given the observed but incomplete data matrix $\mathbf{Y}$ and the current parameter estimates Ω, the complete data log-likelihood is:

$$logL_C(\Omega, \boldsymbol{Y}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{r=1}^{R} z_{ir} y_{ij} log(\theta_{rj}) + (1-y_{ij}) log(1-\theta_{rj}) + \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} log(\pi_r)$$

where $\Omega = \mu, \alpha_r$

A priori z_{ir} follows a multinomial distribution:

$$z_i = (z_{i1}, \dots, z_{iR})^T \sim Multinomial(1; \pi_1, \dots, \pi_R), \qquad \sum_{r=1}^{R} z_r = 1.$$

Then a posteriori x_{jc} also follows a multinomial distribution:

$$z_i = (x_{i1}, \dots, z_{iR})^T \sim Multinomial(1; \hat{z}_1, \dots, \hat{z}_R)$$

In **E-step**, estimate the expected value of z_{ir}, $E[z_{ir}] = \hat{z}_{ir}$. This is also known as the posterior probabilities that row i is a member of row group r. Update $\mathbf{Z}$.

$$\hat{z}_{ir} = \frac{\pi_r \prod_{i=1}^{n} y_{ij} log(\theta_{rj}) + (1-y_{ij}) log(1-\theta_{rj})}{\sum_{g=1}^{C} \kappa_g \prod_{i=1}^{n} y_{ij} log(\theta_{rj}) + (1-y_{ij}) log(1-\theta_{rj})}$$

In **M-step**, numerically maximise the complete log-likelihood. Find $\hat{\pi}_r$ and $\hat{\Omega}$ using the constraint $\sum_{r=1}^{R} \pi_r = 1$,

$$\hat{\pi}_r = \sum_{i=1}^{n} \frac{\hat{z}_{ir}}{n}, \qquad \hat{\Omega} = \underset{\Omega}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{r=1}^{R} \hat{z}_{ir} y_{ij} log(\theta_{rj}) + (1-y_{ij}) log(1-\theta_{rj}) \right]$$

B Column clustering for the ordinal models

The overall (incomplete) likelihood is:

$$L(\Omega, \kappa|\mathbf{Y}) = \prod_{j=1}^{m} \sum_{c=1}^{C} \kappa_r \prod_{i=1}^{n} \prod_{k=1}^{q} \hat{\theta}_{ick}^{I(y_{ij=k})}$$

The overall (incomplete) log-likelihood is:

$$logL(\Omega, \kappa|\mathbf{Y}) = \sum_{j=1}^{m} log\left\{ \sum_{c=1}^{C} \left[\kappa_r \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{ick}^{I(y_{ij=k})} \right] \right\}$$

This likelihood may be optimised but as previously mentioned in the row clustering case, it may be slow or difficult, partly because of the summation sign under the log. Introduce a $\boldsymbol{X \times 1}$ missing matrix $\mathbf{X}$ with $\boldsymbol{x_{jc}}$ as the latent column cluster membership for each column where $x_{jc} = 1$ of $j \in c$ and $x_{jc} = 0$ otherwise.

A priori x_{jc} follows a multinomial distribution:

$$x_j = (x_{j1}, \ldots, x_{jC})^T \sim Multinomial(1; \kappa_1, \ldots, \kappa_R), \qquad \sum_{c=1}^{R} x_c = 1.$$

Then a posteriori x_{jc} also follows a multinomial distribution:

$$x_j = (x_{j1}, \ldots, x_{jC})^T \sim Multinomial(1; \hat{x}_1, \ldots, \hat{x}_C)$$

Apply the E-step by estimating the expected value of x_{jc}, $E[x_{jc}] = \hat{x}_{jc}$, also known as the posterior probabilities that column j lies in column group c.

$$\hat{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{ick}^{I(y_{ij=k})}}{\sum_{g=1}^{C} \kappa_g \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{igk}^{I(y_{ij=k})}}$$

Update $\mathbf{X}$. In M-step, differentiating the complete data log-likelihood,

$$logL_C(\Omega, \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{c=1}^{C} \sum_{k=1}^{q} x_{jc} I(y_{ij} = k) log(\theta_{ick}) + \sum_{j=1}^{m} \sum_{c=1}^{C} x_{jc} log(\kappa_c)$$

and using the constraint $\sum_{c=1}^{C} \kappa_c = 1$ to get

$$\hat{\kappa}_c = \sum_{j=1}^{m} \frac{x_{jc}}{m}$$

. The partial derivatives with respect to Ω can also be obtained in this step.

$$\hat{\Omega} = \underset{\Omega}{\text{argmax}} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{c=1}^{C} \hat{x}_{jc} I(y_{ij} = k) log(\hat{\theta}_{ick}) \right]$$

This two-step iteration repeats until parameter estimates have converged, that is until there is a small relative change in the parameters.

C Column clustering for the binary models

The overall (incomplete) likelihood is:

$$L(\Omega, \kappa | \mathbf{Y}) = \prod_{j=1}^{m} \sum_{c=1}^{C} \kappa_r \prod_{i=1}^{n} \theta_{ic}^{y_{ij}} (1 - \theta_{ic})^{1-y_{ij}}$$

The overall (incomplete) log-likelihood is:

$$logL(\Omega, \kappa | \mathbf{Y}) = \sum_{j=1}^{m} log \left\{ \sum_{c=1}^{C} \left[\kappa_r \prod_{i=1}^{n} \theta_{ic}^{y_{ij}} (1 - \theta_{ic})^{1-y_{ij}} \right] \right\}$$

This likelihood may be optimised but as previously mentioned in the row clustering case, it may be slow or difficult, partly because of the summation sign under the log. Introduce a $\boldsymbol{X} \times \mathbf{1}$ missing matrix $\mathbf{X}$ with $\boldsymbol{x_{jc}}$ as the latent column cluster membership for each column where $x_{jc} = 1$ of $j \in c$ and $x_{jc} = 0$ otherwise.

A priori x_{jc} follows a multinomial distribution:

$$x_j = (x_{j1}, \ldots, x_{jC})^T \sim Multinomial(1; \kappa_1, \ldots, \kappa_R), \qquad \sum_{c=1}^{R} x_c = 1.$$

Then a posteriori x_{jc} also follows a multinomial distribution:

$$x_j = (x_{j1}, \ldots, x_{jC})^T \sim Multinomial(1; \hat{x}_1, \ldots, \hat{x}_C)$$

Apply the E-step by estimating the expected value of x_{jc}, $E[x_{jc}] = \hat{x}_{jc}$, also known as the posterior probabilities that column j lies in column group c.

$$\hat{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n} y_{ij} log(\theta_{ic}) + (1 - y_{ij}) log(1 - \theta_{ic})}{\sum_{g=1}^{C} \kappa_g \prod_{i=1}^{n} y_{ij} log(\theta_{ig}) + (1 - y_{ij}) log(1 - \theta_{ig})}$$

Update $\mathbf{X}$. In M-step, differentiating the complete data log-likelihood,

$$logL_C(\Omega, \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{c=1}^{C} x_{jc} y_{ij} log(\theta_{ic}) + (1 - y_{ij}) log(1 - \theta_{ic}) + \sum_{j=1}^{m} \sum_{c=1}^{C} x_{jc} log(\kappa_c)$$

and using the constraint $\sum_{c=1}^{C}\kappa_c = 1$ to get $\hat{\kappa}_c = \sum_{j=1}^{m}\frac{x_{jc}}{m}$. The partial derivatives with respect to Ω can also be obtained in this step.

$$\hat{\Omega} = \underset{\Omega}{\text{argmax}}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{c=1}^{C}\hat{x}_{jc}y_{ij}log(\theta_{ic}) + (1-y_{ij})log(1-\theta_{ic})\right]$$

This two-step iteration repeats until parameter estimates have converged, that is until there is a small relative change in the parameters.

D Biclustering for the binary models

Introduce a $\boldsymbol{X \times 1}$ missing matrix $\mathbf{X}$ with $\boldsymbol{x_{jc}}$ as the latent column cluster membership for each column where $x_{jc} = 1$ of $j \in c$ and $x_{jc} = 0$ otherwise. The incomplete-data likelihood is:

$$L(\Omega, \pi, \kappa|\mathbf{Y}) = \sum_{c_1=1}^{C}\cdots\sum_{c_m=1}^{C}\kappa_{c_1}\ldots\kappa_{c_m}\sum_{r_1=1}^{R}\cdots\sum_{r_n=1}^{R}\pi_{r_1}\ldots\pi_{r_n}\prod_{i=1}^{n}\prod_{j=1}^{m}\theta_{r_ic_j}^{y_{ij}}(1-\theta_{r_ic_j})^{1-y_{ij}}$$

Assuming row-based conditional independence and independence over the columns, the incomplete-data log-likelihood is given as:

$$logL(\Omega, \pi, \kappa|\mathbf{Y}) = log\left\{\sum_{c_1=1}^{C}\cdots\sum_{c_m=1}^{C}\kappa_{c_1}...\kappa_{c_m}\prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r\prod_{j=1}^{m}\theta_{r_ic_j}^{y_{ij}}(1-\theta_{r_ic_j})^{1-y_{ij}}\right]\right\}$$

Consequently, the complete data log-likelihood is:

$$\ell_C(\Omega, Y, Z, X) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{r=1}^{R}\sum_{k=1}^{q}z_{ir}x_{jc}y_{ij}log(\theta_{rc}) + (1-y_{ij})log(1-\theta_{rc}) + \sum_{i=1}^{n}\sum_{r=1}^{R}z_{ig}log(\pi_r) + \sum_{j=1}^{m}\sum_{c=1}^{C}x_{jc}log(\kappa_c)$$

In **E-step**, given $\mathbf{Y}$, the missing matrices $\mathbf{Z}$ and $\mathbf{X}$, and some chosen values for starting points Ω, the expected values of z_{ir} and x_{jc} are as follows,

$$\hat{z}_{ir} = \frac{\pi_r\prod_{j=1}^{m}\{\sum_{c=1}^{C}\kappa_c y_{ij}log(\theta_{rc}) + (1-y_{ij})log(1-\theta_{rc})\}}{\sum_{b=1}^{C}\kappa_c\prod_{i=1}^{n}\{\sum_{a=1}^{R}\pi_a y_{ij}log(\theta_{ab}) + (1-y_{ij})log(1-\theta_{ab})\}},$$

$$\hat{x}_{jc} = \frac{\kappa_c\prod_{i=1}^{n}\{\sum_{r=1}^{R}\pi_r y_{ij}log(\theta_{rc}) + (1-y_{ij})log(1-\theta_{rc})\}}{\sum_{b=1}^{C}\kappa_c\prod_{i=1}^{n}\{\sum_{a=1}^{R}\pi_a\prod_{i=1}^{n}y_{ij}log(\theta_{ab}) + (1-y_{ij})log(1-\theta_{ab})\}}$$

Update $\mathbf{Z}$ and $\mathbf{X}$. In **M-step**, numerically maximising the complete-data log-likelihood to obtain the maximum likelihood estimates $\hat{\Omega}$, $\hat{\pi}_r$ and $\hat{\kappa}_c$,

$$\hat{\Omega} = \underset{\Omega}{\text{argmax}}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{r=1}^{R}\sum_{c=1}^{C}\hat{z}_{ir}\hat{x}_{jc}y_{ij}log(\theta_{rc}) + (1-y_{ij})log(1-\theta_{rc})\right], \quad \hat{\pi}_r = \sum_{i=1}^{n}\frac{\hat{z}_{ir}}{n}, \quad \hat{\kappa}_c = \sum_{j=1}^{m}\frac{\hat{x}_{jc}}{m}$$

References

Sylvia Fruhwirth-Schnatter, Gilles Celeux, and Christian P.Robert. *Handbook of Mixture Analysis*. Chapman & Hall/CRC, 2019.

Daniel Fernández, Richard Arnold, Shirly Pledger, Ivy Liu, and Roy Costilla. Finite mixture biclustering of discrete type multivariate data. *Advances in Data Analysis and Classification*, 2018. doi: 10.1007/s11634-018-0324-3.

Alan Agresti. *Analysis of Ordinal Categorical Data, Second Edition*. Wiley, 2010. doi: 10.1002/9780470594001.

Shirley Pledger and Richard Arnold. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics & Data Analysis*, 71:241–261, 2014.

A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1977.

Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and Extensions*. Wiley, 2007. ISBN 978-0-470-19160-6.

Bernd Kortmann and Kerstin Kerstin Lunkenheimer, editors. *eWAVE*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://ewave-atlas.org/`.

Ivy Liu, Richard Arnold, Miriam Meyerhoff, Shirley Pledger, and Lingyu Li. *Clustering Language Features: Scaling up from Micro to Macro Variation*. JSM 2019 Online Program, Social Statistics Section, 2019. URL `https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=300030`.

Eleni Matechou, Ivy Liu, Daniel Fernández, Miguel Farias, and Bergljot Gjelsvik. Biclustering models for two-mode ordinal data. *Psychometrika*, 81(3):611–624, 2016.

Roy Costilla. *Clustering repeated ordinal data: Model based approaches using finite mixtures*. PhD thesis, Victoria University of Wellington, 06 2017.

Daniel Fernández. *Mixture-based clustering for the ordered stereotype model*. PhD thesis, Victoria University of Wellington, 2015.

Daniel Fernández, Richard Arnold, and Shirley Pledger. Mixture-based clustering for the ordered stereotype model. *Computational Statistics & Data Analysis*, 93:46–75, 2016.

Daniel Fernández and Shirley Pledger. Categorising count data into ordinal responses with application to ecological communities. *Agricultural, Biological and Environmental Statistics*, 21:348–362, 2015. doi: 10.1007/s13253-015-0240-3.

Daniel Fernández, Roy Costilla, and Ivy Liu. A method for ordinal outcomes: The ordered stereotype model. *Psychiatric Research*, 8, 2019. doi: https://doi.org/10.1002/mpr.1801.

McMillan, Louise and Fernández, Daniel and Matechou, Eleni. *Likelihood-based clustering*. School of Mathematics, Statistics and Operations Research, VUW, 2020. URL `https://github.com/vuw-clustering/clustord.git`.

Daniel Fernández, Richard Arnold, and Shirley Pledger. Introducing spaced mosaic plots. 2014.