Mixture-based Clustering for the Ordered Stereotype Model

by

Daniel Fernández Martínez

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy in Statistics.

Victoria University of Wellington
2015

Abstract

Many of the methods which deal with the reduction of dimensionality in matrices of data are based on mathematical techniques. In general, it is not possible to use statistical inferences or select the appropriateness of a model via information criteria with these techniques because there is no underlying probability model. Furthermore, the use of ordinal data is very common (e.g. Likert or Braun-Blanquet scale) and the clustering methods in common use treat ordered categorical variables as nominal or continuous rather than as true ordinal data. Recently a group of likelihood-based finite mixture models for binary or count data has been developed (Pledger and Arnold, 2014). This thesis extends this idea and establishes novel likelihood-based multivariate methods for data reduction of a matrix containing ordinal data. This new approach applies fuzzy clustering via finite mixtures to the ordered stereotype model (Fernández et al., 2014a). Fuzzy allocation of rows and columns to corresponding clusters is achieved by performing the EM algorithm, and also Bayesian model fitting is obtained by performing a reversible jump MCMC sampler. Their performances for one-dimensional clustering are compared. Simulation studies and three real data sets are used to illustrate the application of these approaches and also to present novel data visualisation tools for depicting the fuzziness of the clustering results for ordinal data. Additionally, a simulation study is set up to empirically establish a relationship between our likelihood-based methodology and the performance of eleven information criteria in common use. Finally, clustering comparisons between count data and categorising the data as ordinal over a same data set are performed and results are analysed and presented.

Acknowledgments

I would like to thank all the people who contributed in some way to the work described in this thesis. First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Richard Arnold whose immense expertise, enthusiasm, encouragement, supervision, and patience have been priceless. His academic guidance has been invaluable, helping me from the very early stage to the conclusion of this thesis.

I would like to show my gratitude to Prof. Shirley Pledger, my co-supervisor, for her supervision, advice, and persistent help. I am very grateful to her for supporting me both professionally and personally throughout my study.

It has been an extraordinary experience to work with Dr. Arnold and Prof. Pledger and this thesis would not have been possible without their supervision. I am specially grateful to them for giving me intellectual freedom in my work, supporting my attendance at various conferences, engaging me in new ideas, and helping me in the revision of this thesis.

I would also like to make a special reference to Dr. Ivy (I Ming) Liu who has always been available to answer my questions. Her vast expertise in categorical data has been a crucial contribution to this thesis.

I would like to express my very great appreciation to the three examiners, Prof. Geoff McLachlan, Prof. Richard Barker and Dr. Yuichi Hirose, for their encouragement and their critical comments about this thesis.

I would like to acknowledge Victoria University of Wellington (VUW) for providing me with financial support through Victoria Doctoral Scholarship program, and giving me the opportunity to work as a teaching assistant and tutor. I gratefully thank to all the academics and staff from the School of Mathematics Statistics and Operations Research (MSOR) at VUW to support me when I needed it. In particular, I would like to reference to Dr. Petros Hadjicostas for his suggestions on mathematical aspects of my thesis. I would also like to acknowledge all my

officemates in CO528 who kept me in good company during all this time.

I gratefully acknowledge Dr. Mark Tozer from Science Division, NSW Office of Environment and Heritage in Australia and Dr. Royce Anders from University of California, Irvine for providing some of the data sets I have used in this thesis. I also thank Dr. Murray Efford for his guidance on the creation of R packages.

I am especially thankful for the assistance given by my dear friend Corinna Howland for proofreading this thesis but I also thank her for her friendship and unyielding support.

A special thanks to my dear friends Andrea Kolb, Holly O'Connor, Mario Alayón, Paula Ryan, Ricky Ferrando, Roz Walls, Thanh Nguyen, Valentina Baccetti, You Wang Qing (Diana), David Peña, José Luís Estévez, Klaus Langohr, Lesly Acosta, Marc Martínez, Víctor García, and all my friends around the world who have always believed that I could do it.

I owe my deepest gratitude to my mother who gave me the opportunity of an education and has supported me throughout my life, and also to my brother who has always been a big supporter of this adventure in New Zealand.

Last but not least, I would like to thank the love of my life Pilar. She has been my support every single moment during the process of this thesis. I am very lucky to have her in my life and cannot imagine it without her.
To her and to my baby-to-come Paola I dedicate this thesis.

The Cuban poet José Martí once said: "Every person should plant a tree, have a child, and write a book. These all live on after us, insuring a measure of immortality".
Here is my book.

Contents

Chapter 1

Introduction

1.1 Ordinal Data

An ordinal variable is one with a categorical data scale which describes order, and where the distinct levels of such a variable differ in degree of dissimilarity more than in quality (Agresti, 2010). This is different from nominal variables which vary in quality, not in quantity, and thus the order of listing the categories is irrelevant. In his seminal paper, Stevens (1946) called a scale ordinal if "any order-preserving transformation will leave the scale form invariant". Examples of ordinal variables are the measures of the effectiveness of a new drug ("low", "medium" or "high"), the pain scale (see Figure 1.1), the Likert scale responses in a questionnaire might be "disagree", "neither agree nor disagree" or "agree", or the cover-abundance scale of Braun-Blanquet or Domin in vegetation science. An

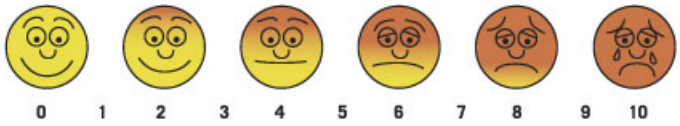

Figure 1.1: ***Pain scale****: This scale measures a patient's pain intensity in which 0 means no pain and 10 means extremely painful. Pain scales are based on self-report, observational, or physiological data.*

important point to notice is the degree of dissimilarity among the different levels of the scale in an ordinal variable might not necessarily be always the same. For instance, the difference in the severity of an injury expressed by level 2 rather than level 1 might be much more than the difference expressed by a rating of level 10

rather than 9. In addition, the orientation of the ordered categories (from high to low or from low to high) is not relevant to the conclusions over the ordinal data. However, the way the categories are ordered in the data is relevant as it could change the results of the analysis.

Although the collection and use of ordinal variables is common, most of the current methods for analysing them treat the data as if they were nominal (Hoffman and Franke, 1986) or continuous data (Agresti, 2010). On the one hand, treating an ordered categorical variable as ordinal rather than nominal provides advantages in the analysis such as simplifying the data description and allowing the use of more parsimonious models. The nominal approach ignores the intrinsic ordering of the data and thus the statistical results are less powerful than they could be. On the other hand, models for continuous variables have similarities to those for ordinal variables although the use of them with ordinal variables has disadvantages such as the treatment of the output categories as equally spaced, which they may not be (see Agresti (2010, Sections 1.2-1.3) for a list of advantages of treating an ordinal variable as ordinal rather than nominal or continuous).

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

1.2 Clustering of Categorical Response Data

Nowadays, many studies from different disciplines are related to variables which are measured in subjects which are classified in clusters (see Figure 1.2). Basically, the composition of each cluster is determined by the degree of similarity among subjects or on a set of repeated measures of the same subject through a specific period of time (e.g. longitudinal studies) or space (e.g. community ecology data). In the research literature, many algorithms and techniques have been developed

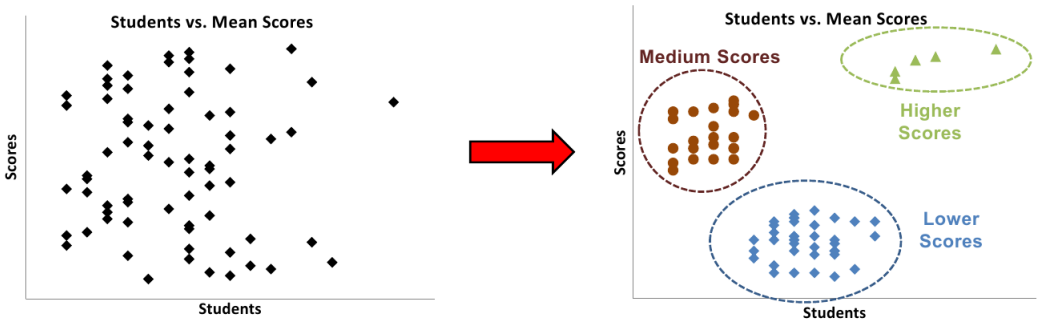

Figure 1.2: ***Clustering**: Measures of the scores of 70 students in a particular subject. The raw data is shown on the left graph. After rearranging the data, three clusters are identified on the right graph (students with lower, medium and higher scores).*

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

et al. (2008) and Melnykov and Maitra (2010). We describe more details of this approach below (Section 1.2.1).

The simultaneous clustering of rows and columns into row clusters and column clusters is called biclustering (or block clustering, two-dimensional clustering or two-mode clustering). Biclustering models based on double k-means have been developed in Vichi (2001) and Rocci and Vichi (2008). A hierarchical Bayesian procedure for biclustering is given in DeSarbo et al. (2004). Biclustering using mixtures has been proposed for binary data in Pledger (2000), Arnold et al. (2010) and Labiod and Nadif (2011), and for count data in Govaert and Nadif (2010). An approach via finite mixtures for binary and count data using basic Bernoulli or Poisson building blocks has been developed in Govaert and Nadif (2010) and Pledger and Arnold (2014). This work expanded previous research for one-mode fuzzy cluster analysis based on finite mixtures to a suite of models including biclustering. Finally, Matechou et al. (2011) have recently developed biclustering models for ordinal data using the assumption of proportional odds and having a likelihood-based foundation. The main difference with our work is that we use the assumption of ordinal stereotype model which has the advantage of allowing us to determine a new spacing of the ordinal categories, dictated by the data. We develop the formulation of this model in Chapter 2.

1.2.1 Clustering Analysis Based on Finite Mixtures

The widespread use of finite mixture models as a mathematical-based method for statistical modeling of unknown random phenomena in an extremely flexible way has increased over the last 20 years (McLachlan and Peel, 2000). An appropriate choice of the components that make up the finite mixture model allows both the accurate representation of complex distributions and the inference about the random phenomena observed. In addition, the application of finite mixture models is useful in a variety of statistical techniques such as clustering, discriminant analysis and image analysis where the main motivation is the modelling of heterogeneity through the identification of different groups or classes. Further advantages of finite mixtures modeling in comparison with other clustering methods include the better handling of missing data and the possibility to fit structured data (e.g. longitudinal data).

Finite mixture modeling can be viewed as a latent variable analysis with a la-

tent categorical variable describing the group or subpopulation membership and the latent classes being described by the different components of the mixture density (Skrondal and Rabe-Hesketh, 2004). The usefulness of finite mixture models as a strategy for doing data clustering and, in that way, exposing the distinct classes that may underlie the data was first explained in McLachlan and Basford (1988). This approach is called the mixture model-based approach to clustering and it is assumed that the data come from a mixture of a specified number of groups R, where each observation is a realization y from the following finite mixture density,

$$f(y;\Omega) = \sum_{r=1}^{R} \pi_r f_r(y;\theta_r).$$

Here Ω contains all the unknown parameters in the mixture, θ_r is the vector of unknown parameters in the rth component density of the finite mixture $f_r(y;\theta_r)$ and $\pi_1,\ldots,\pi_R$ are nonnegative quantities where

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

and represent the probability of being a member of the group r.

The fitting of this model can be done by maximum likelihood (ML) estimation. Bayesian approaches to achieve this estimation can also be used. In ML estimation the optimisation of the likelihood is simplified considerably by using the iterative expectation-maximisation (EM) algorithm considering group membership as missing data (Dempster et al., 1977). The problem therefore becomes a classical case of ML estimation from data that can be viewed as being incomplete. It is important to mention there are two possible issues to take into account using the EM algorithm for fitting in a finite mixture model context. Firstly, it is quite common to find multimodality of the likelihood when we deal with this type of model. A related issue is how to select the suitable starting parameter values for the EM algorithm since different starting values will lead the algorithm to different local maxima. A recommended strategy is to employ an iterative framework where several starting values are tested over the parameter space.

An important consideration with this fitting concerns the choice of the number of components R in the finite mixture. It is common to use information criteria such as AIC (Akaike, 1973), AICc (Hurvich and Tsai, 1989), BIC (Schwarz, 1978)

and ICL.BIC (Biernacki et al., 1998) in order to select a suitable number of groups R. Additionally, it is not possible to use likelihood ratio test (LRT) as a model selection procedure because the regularity conditions do not hold for the log-likelihood ratio statistic $-2\mathrm{LLR}$ to have asymptotic null distribution in mixture densities framework, where LLR is the test statistic defined in eq. (3.1). However, the use of the LRT is still possible when the null distribution is assessed, for example, by a bootstrap approach (McLachlan and Peel, 2000, Section 6.4-6.5).

Once the finite mixture model has been fitted, the estimated prior probabilities (mixing proportions) are obtained, $\hat{\pi}_1, \ldots, \hat{\pi}_R$. In addition, the probabilistic (fuzzy) clustering is given in terms of the estimated posterior probabilities of component membership for observation i being classified into each group, $(\hat{Z}_{i1}, \ldots, \hat{Z}_{iR})$. These are defined a priori as

$$(Z_{i1}, \ldots, Z_{iR}) \sim \mathrm{Multinomial}(1; \pi_1, \ldots, \pi_R),$$

and a posteriori, conditional on the data Y, as

$$(Z_{i1}, \ldots, Z_{iR})|Y \sim \mathrm{Multinomial}(1; \hat{Z}_{i1}, \ldots, \hat{Z}_{iR}).$$

We note that the sample size is 1 as $\sum_{r=1}^{R} Z_{ir} = 1$. For that reason, one possible way to assign each realization to one single component, if that were needed, would be allocating the observation to the cluster with highest marginal posterior probability of belonging. Alternatively, randomly assign each observation by drawing from the expected value of the posterior probabilities $Z_{i1}, \ldots, Z_{iR}$, conditional on the data.

1.3 Ordinal Modelling

Throughout this thesis the ordinal responses Y are labeled $1, 2, \ldots, q$. There are a variety of approaches to the modelling of ordinal data. We will employ methods which properly respect the ordinal nature of the data, without the assumption that the data are continuous. Among ordinal models there are a variety of modelling strategies. This thesis is focused on the *ordered stereotype model* (Anderson, 1984) which is thoroughly defined in Chapter 2. A brief review of some other common logistic regression models for ordinal response variables follows (see a

full review in Agresti (2010, Chapters 3 and 4)):

The Proportional Odds Model Version of the Cumulative Logit Model

The proportional odds model is one of the most popular models for ordinal responses, and it became popular after the article of McCullagh (1980). In order to describe this model, first we define the *cumulative logits* for a q-category ordinal response variable Y as follows,

$$\begin{aligned}\text{logit}\left[P\left(Y \leq k \mid \boldsymbol{x}\right)\right] &= \log\left(\frac{P\left(Y \leq k \mid \boldsymbol{x}\right)}{1 - P\left(Y \leq k \mid \boldsymbol{x}\right)}\right) \\ &= \log\left(\frac{p_1 + \ldots + p_k}{p_{k+1} + \ldots + p_q}\right), \qquad k = 1, \ldots, q-1,\end{aligned}$$

where $p_1, \ldots, p_q$ are denoting the response probabilities, which satisfy $\sum_{k=1}^{q} p_k = 1$, and $\boldsymbol{x}$ represents a set of predictor variables which can be quantitative or categorical (with indicator variables). It is important to note that each cumulative logit depends on all q response categories. We also note that this model treats each logit as a model for a binary variable, in which the first collapsed k categories form one of the outcomes and the categories from $k+1$ to q form the other one.

We define the *proportional odds model version of the cumulative logit* model by using all q cumulative logits simultaneously,

$$\text{logit}\left[P\left(Y \leq k \mid \boldsymbol{x}\right)\right] = \mu_k - \boldsymbol{\delta}'\boldsymbol{x}, \qquad k = 1, \ldots, q-1, \tag{1.1}$$

with the monotone increasing ordinal constraint $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{q-1}$. Each cumulative logit has its own intercept $(\mu_1, \ldots, \mu_{q-1})$ called its *cut point*. These parameters are generally of little interest, and we might usually consider them as nuisance parameters. The parameter vector $\boldsymbol{\delta}$, which describes the effect of $\boldsymbol{x}$ on the log odds of the response variable in the category k or below, is independent of that category k. This independence is the proportional odds assumption. Therefore, all the cumulative logits contain the same effect of $\boldsymbol{x}$. This has the advantage that we only have to fit one parameter instead of $q-1$ parameters. In this manner, this model assumes that all the effects $\boldsymbol{\delta}$ over a set of predictor variables $\boldsymbol{x}$ on the response variable Y for the defined log odds are the same regardless of any collapsing of the q-level response variable to a binary variable.

The degree of association of the response Y with the predictor variables $\boldsymbol{x}$ is determined by the value of the effects $\boldsymbol{\delta}$. When the model fitted shows that $\boldsymbol{\delta} = 0$, the value of $P(Y \leq k)$ as a function of $\boldsymbol{x}$ is a constant for each k and it means that Y and $\boldsymbol{x}$ are statistically independent. The negative sign preceding the effect $\boldsymbol{\delta}$ for predictor $\boldsymbol{x}$ in expression (1.1) allows a parametrisation which has a natural interpretation of the effect $\boldsymbol{\delta}$ regarding whether it is positive or negative. Therefore, if $\boldsymbol{\delta} > 0$ then higher values of $\boldsymbol{x}$ lead to higher values of Y. This interpretation is reversed when $\boldsymbol{\delta} < 0$.

The *proportional odds model version of the cumulative logit* model satisfies the *proportional odds* property. It is described as follows,

$$\begin{aligned} &\operatorname{logit}\left[P\left(Y \leq k \mid \boldsymbol{x}_1\right)\right] - \operatorname{logit}\left[P\left(Y \leq k \mid \boldsymbol{x}_2\right)\right] \\ &= \log\left[\frac{P\left(Y \leq k \mid \boldsymbol{x}_1\right)/P\left(Y > k \mid \boldsymbol{x}_1\right)}{P\left(Y \leq k \mid \boldsymbol{x}_2\right)/P\left(Y > k \mid \boldsymbol{x}_2\right)}\right] = \boldsymbol{\delta}'(\boldsymbol{x_1} - \boldsymbol{x_2}). \end{aligned} \tag{1.2}$$

An odds ratio of cumulative probabilities is called a *cumulative odds ratio*. The logarithm of a cumulative odds ratio is the difference between the cumulative logits at $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ and it is proportional to the distance between them and independent of k. In addition, the odds of making response $Y \leq k$ at $\boldsymbol{x} = \boldsymbol{x_1}$ are $\exp\left[\boldsymbol{\delta'}(\boldsymbol{x_1} - \boldsymbol{x_2})\right]$ times the odds at $\boldsymbol{x} = \boldsymbol{x_2}$.

Adjacent-Categories Logit Models

The definition for all the pairs of adjacent categories logits is as follows,

$$\log\left(\frac{p_k}{p_{k+1}}\right), \qquad k = 1, \ldots, q-1,$$

where $(p_1, ..., p_q)$ are the set of probabilities regarding to the q outcome categories. In this manner, we redefine the *adjacent-categories logit* model as follows,

$$\log\left(\frac{p_k(\boldsymbol{x})}{p_{k+1}(\boldsymbol{x})}\right) = \mu_k - \boldsymbol{\delta}_k'\boldsymbol{x}, \qquad k = 1, \ldots, q-1. \tag{1.3}$$

As we observed, it takes into account the probability of two adjacent-categories rather than considering the probability of each category versus a baseline-category. For this reason, the effects $\boldsymbol{\delta}_k$ are described with local log odds instead of the cumulative log odds. The adjacent structure of this model recognizes the ordering

of categories of the response variable Y (Agresti, 2010, Section 4.1.1). Thus, we can apply the proportional odds property to the linear predictor in order to obtain a more parsimonious version of the model, in the case where an explanatory variable has a similar effect for all the logits. A version of the adjacent-categories logit with proportional odds model for the baseline-category is

$$\log\left(\frac{P\left[Y=k \mid \boldsymbol{x}\right]}{P\left[Y=1 \mid \boldsymbol{x}\right]}\right)=\mu_k-\boldsymbol{\delta}'\boldsymbol{x}, \qquad k=1,\ldots,q-1. \tag{1.4}$$

This model is called *standard baseline-category logit with proportional odds* model. The difference is that the logits are not defined by adjacency but always with respect to a single baseline category.

Continuation-Ratio Logit Models

There are two types of settings for the *continuation-ratio logit* model. One is based on the log odds of each category related to the lower categories of the variable response,

$$\log\left(P[Y=k \mid Y\leq k]\right)=\log\left(\frac{p_k}{p_1+\ldots+p_k}\right), \qquad k=1,\ldots,q-1,$$

and the other one is based on the log odds related to the higher categories,

$$\log\left(P[Y=k \mid Y\geq k]\right)=\log\left(\frac{p_k}{p_k+\ldots+p_q}\right), \qquad k=1,\ldots,q-1.$$

This last variety of log odds is extensively used in the case of analysis where all the categories define states with an established order where the subject passes through the different categories before the response outcome is determined, e.g. recovery time of subjects who are observed after a cancer treatment (less than 1 year, 1 to 2 years, 2 to 3 years, 3 to 5 years, more than 5 years). This type of process is called a *sequential process*. Finally, according to this and using the set of predictors $\boldsymbol{x}$, we can describe the *continuation-ratio logit with proportional odds model for a sequential process* (McCullagh and Nelder, 1989) as follows,

$$\operatorname{logit}\left(P[Y=k \mid Y\geq k]\right)=\operatorname{logit}\left[\frac{p_k}{p_k+\ldots+p_q}\right]=\mu_k-\boldsymbol{\delta}'\boldsymbol{x}, \qquad k=1,\ldots,q-1.$$

Additionally, it is possible to describe this model in its partial proportional odds structure (see Cole and Ananth (2001)).

Multinomial Logistic Regression

Considering the first category as the baseline category, the model is defined with the following $q-1$ simultaneous log odds,

$$\log\left(\frac{P\left[Y=k \mid \boldsymbol{x}\right]}{P\left[Y=1 \mid \boldsymbol{x}\right]}\right)=\mu_k+\boldsymbol{\delta}_k'\boldsymbol{x}, \qquad k=2,\ldots,q, \tag{1.5}$$

where Y is the response variable which has q categories, $\boldsymbol{x}$ is the vector of predictor variables which can be categorical or continuous variables, $\{\mu_2 \ldots \mu_q\}$ are the intercept parameters for each category and $\boldsymbol{\delta}_k$ represents the vector of parameters of the effects of $\boldsymbol{x}$ on the log odds of the response variable for the category k related to the baseline category. In order to identify the model, we need to place constraints on the parameters. Commonly μ_1 and δ_1 are constrained to be 0.

Table 1.1: Degree of Suffering from Disturbed Dreams, by age.

	Degree of Suffering			
Age	Not Severe	Low Severe	Medium Severe	Very Severe
5-7	7	4	3	7
8-9	10	15	11	13
10-11	23	9	11	7
12-13	28	9	12	10
14-15	32	5	4	3

Source: Data from Maxwell (1961).

The model assumes that different linear combinations of the predictor variables are required in order to discriminate between the $q-1$ pairs of log odds of the response variable. Unlike the proportional odds version of the cumulative logit model (1.1), the model formulated as (1.5) is not specifically embodied in ordinal response variables. However, we present it here with the aim of helping us in the interpretation of the ordered stereotype model in Chapter 2. Thus, this model has $q-1$ log odds, one for each comparison between the $q-1$ categories and the baseline category. Therefore, each element of the predictor vector $\boldsymbol{x}$ has $q-1$ different parameters. In order to illustrate this, we use an example of the

severity of disturbed dreams among boys between 5 and 15 years old (see Table 1.1) which has $q = 4$ categories and suppose that there are three predictor variables for each boy. Thus, the multinomial logistic regression model is defined with the following pairs of log odds,

$$\log\left(\frac{P\left[Y=2 \mid \boldsymbol{x}\right]}{P\left[Y=1 \mid \boldsymbol{x}\right]}\right) = \mu_2 + \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3,$$

$$\log\left(\frac{P\left[Y=3 \mid \boldsymbol{x}\right]}{P\left[Y=1 \mid \boldsymbol{x}\right]}\right) = \mu_3 + \delta_{31}x_1 + \delta_{32}x_2 + \delta_{33}x_3, \text{ and} \tag{1.6}$$

$$\log\left(\frac{P\left[Y=4 \mid \boldsymbol{x}\right]}{P\left[Y=1 \mid \boldsymbol{x}\right]}\right) = \mu_4 + \delta_{41}x_1 + \delta_{42}x_2 + \delta_{43}x_3,$$

which has 12 parameters to estimate. For that reason, the model is not parsimonious and if the number of categories q or the number of predictor variables is large, the model might over-parametrise the data and, therefore, the interpretation of the results is difficult.

1.4 Outline of the Thesis

This thesis presents an extension of the likelihood-based models proposed in Pledger and Arnold (2014) consisting in applying them to matrices with ordinal data via finite mixtures to define a fuzzy clustering. We use the ordered stereotype model introduced by Anderson (1984) to formulate the ordinal approach. The methodology, model fitting and data applications for our approach are separately presented in different chapters.

Chapter 2 is a review of the stereotype model that includes its definition and interpretation. Additionally, a formulation of this model including fuzzy clustering via finite mixtures is shown. We review the methodologies proposed in the literature to fit this model. Model fitting for the clustering version by using the iterative EM algorithm is described in this chapter .

There have been several approaches via finite mixture models to solve the classification problem of deciding how many clusters are in a given data set. Chapter 3 presents a review of several model comparison measures. In order to

test which measure is the most reliable for our ordinal approach, we set up simulation studies to compare the performance of eleven information criteria. Their results are shown in this chapter. Moreover, the best information criteria from the experimental study are tested applying our approach to a real-life dataset which is intentionally collected to be composed of a known number of clusters.

The reliability of estimation of the stereotype model parameters is demonstrated in a simulation study in Chapter 4. Additionally, we illustrate the application of our likelihood-based finite mixture model method with three real-life examples. Model comparison is applied to these examples based on the results obtained in Chapter 3.

There are a number of visualisation tools that can help to depict the reduction of dimensionality in matrices of ordinal data such as multidimensional scaling and correspondence analysis plots. Chapter 5 introduces new graphical tools for ordinal data based on mosaic, level and contour plots. Furthermore, the R function we developed including some of these novel graphs is introduced in this chapter.

Chapter 6 presents a comparison of clustering results between count and categorised ordinal data. A review of a stochastic scheme for classifying count data in relation to its variance-mean ratio is shown and some advantages of categorising count data into ordinal are enumerated. In addition, a strategy for determining the optimal number of ordinal categories is presented. Clusterings from count and ordinal data methods are compared by using three measures over the same data set: the adjusted Rand index, the normalized variation of information and the normalized information distance. These comparison measures are described in this chapter.

Chapters 7 and 8 introduce a Bayesian approach to parameter estimation for our ordinal clustering procedure. Chapter 7 enumerates some key factors to consider in Markov chain Monte Carlo (MCMC) samplers in order to assess its reliability and convergence diagnostics. The framework to implement the Metropolis-Hastings sampler for our method is developed and illustrated with a simulation study and two real-life data examples. Additionally, the label switching problem, which is a common drawback arising from using mixture models, is described. Chapter 8 shows the development of a reversible jump MCMC (RJMCMC) sampler which estimates the number of clusters and parameters simultaneously from their joint posterior distribution for our clustering approach. In addition, the con-

vergence diagnostic for RJMCMC samplers is presented and the application of the sampler is illustrated with a simulation study and two real-life data examples.

We conclude with final remarks and discussion in Chapter 9.

All the programs throughout this thesis are written in **R** (Development Core Team (2010)) with certain functions compiled in **C** code called from **R** for the sake of computational speed.

Chapter 2

Ordered Stereotype Model

2.1 Introduction

2.1.1 Data and Model Definition

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

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

The *ordered stereotype model* was introduced by Anderson (1984) and is a model to analyse categorical response variables (see the description of other models in Section 1.3). This model for the probability that y_{ij} takes the category k is characterized by the following log odds

$$\log\left(\frac{P\left[y_{ij} = k \mid \boldsymbol{x}\right]}{P\left[y_{ij} = 1 \mid \boldsymbol{x}\right]}\right) = \mu_k + \phi_k \boldsymbol{\delta'} \boldsymbol{x}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, m, \quad k = 2, \ldots, q, \tag{2.1}$$

where the inclusion of the following monotone increasing constraint

$$0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1 \tag{2.2}$$

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

$$P\left[y_{ij} = k \mid \boldsymbol{x}\right] = \frac{\exp(\mu_k + \phi_k \boldsymbol{\delta'} \boldsymbol{x})}{\sum_{\ell=1}^{q} \exp(\mu_\ell + \phi_\ell \boldsymbol{\delta'} \boldsymbol{x})}, \qquad \text{for } k = 1, \ldots, q, \tag{2.3}$$

where the probability for the baseline category, as defined in (2.3), satisfies

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

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

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

Greenland (1994) showed that the stereotype model is a natural option when the progression of the response variable occurs through various stages and Agresti (2010, Chapter 4) showed that the stereotype model is equivalent to an ordinal model such as the proportional odds version of the adjacent-categories logit model (1.3), when the scores $\{\phi_k\}$ are a linear function of the different categories of the response variable. That is, the stereotype model formulated in (2.1) can be

reformulated in terms of adjacent categories logits (1.3) as follows,

$$\begin{aligned}
\log\left(\frac{P\left[y_{ij}=k \mid \boldsymbol{x}\right]}{P\left[y_{ij}=k+1 \mid \boldsymbol{x}\right]}\right) &= \log\left(\frac{P\left[y_{ij}=k \mid \boldsymbol{x}\right]/P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}{P\left[y_{ij}=k+1 \mid \boldsymbol{x}\right]/P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}\right) \\
&= \log\left(\frac{P\left[y_{ij}=k \mid \boldsymbol{x}\right]}{P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}\right) - \log\left(\frac{P\left[y_{ij}=k+1 \mid \boldsymbol{x}\right]}{P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}\right) \\
&= (\mu_k - \mu_{k+1}) + (\phi_k - \phi_{k+1})\boldsymbol{\delta'x} \\
&= \eta_k + \vartheta_k \boldsymbol{\delta'x}, \qquad\qquad k = 2, \ldots, q,
\end{aligned} \tag{2.4}$$

where $\eta_k = \mu_k - \mu_{k+1}$ ($k = 1, \ldots, q-1$). The relation between $\{\phi_k\}$ and $\{\vartheta_k\}$ is defined by

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

and

$$\phi_k = \sum_{t=1}^{k-1} \vartheta_t, \qquad k = 1, \ldots, q-1.$$

Therefore, the adjacent-categories logit model (see eq. (1.3)) is a particular case of the ordered stereotype model when $\vartheta_k = 1$ in eq. (2.4) or, in other words, when the $\{\phi_k\}$ scores are fixed and equally spaced.

2.1.2 Interpretation

In the stereotype model, the log odds ratio for the increase of a unit of a specific covariate x_ℓ (from u to $u+1$) for a particular categorical response k is as follows,

$$\begin{aligned}
&\log\left(\frac{P\left[y_{ij}=k \mid x_\ell=u+1\right]/P\left[y_{ij}=1 \mid x_\ell=u+1\right]}{P\left[y_{ij}=k \mid x_\ell=u\right]/P\left[y_{ij}=1 \mid x_\ell=u\right]}\right) \\
&= \log\left(\frac{P\left[y_{ij}=k \mid x_\ell=u+1\right]}{P\left[y_{ij}=1 \mid x_\ell=u+1\right]}\right) - \log\left(\frac{P\left[y_{ij}=k \mid x_\ell=u\right]}{P\left[y_{ij}=1 \mid x_\ell=u\right]}\right) \\
&= \mu_k + \phi_k\delta_\ell(u+1) - \mu_k - \phi_k\delta_\ell u = \phi_k\delta_\ell.
\end{aligned}$$

For that reason, the odds ratio for a category k comparing with the baseline when there is a unit increase in x_ℓ is $\exp(\phi_k\delta_\ell)$. In other words, the coefficient $\phi_k\delta_\ell$ in the stereotype model represents the log odds ratio for categories k and the baseline category of the response variable y_{ij} with a unit increase in the predictor variable x_ℓ. For example, in the data regarding the degree of suffering from

disturbed dreams in boys by their age (see Table 1.1 in Chapter 1), the predictor variable is the age of the boy and its estimated parameter is $\hat{\delta}_{age} = 0.31$. Therefore, the estimated odds ratio comparing the "not severe" category (the baseline) versus the "medium severe" category (with estimated score parameter $\hat{\phi}_3 = 0.36$) is $\exp(0.36(0.31)) = 1.12$. In other words, the odds of a boy of suffering medium severe dreams instead of not suffering severe dreams are 1.12 times the odds when the boy is one year younger. Note that the constraint $\phi_q = 1$ implies that the coefficients δ_ℓ, corresponding to x_ℓ, represents the effect of a unit change in x_ℓ on the log odds ratio of response in the highest category q versus the baseline category of y_{ij}.

The order constraint on the scores $\{\phi_k\}$, expressed in (2.2), implies that for a unit increase in the predictor variable x_ℓ, the odds ratio $\exp(\phi_k \delta_\ell)$ of category k vs. baseline category becomes larger when category k is further from the baseline category. One way to interpret how the order constraint on the scores $\{\phi_k\}$ gives an ordinal character to the response variable y_{ij} is by means of the formulation of the stereotype model for two particular categories a and b of the response variable y_{ij}. Thus,

$$\begin{aligned}
\log\left(\frac{P\left[y_{ij}=a \mid \boldsymbol{x}\right]}{P\left[y_{ij}=b \mid \boldsymbol{x}\right]}\right) &= \log\left(\frac{P\left[y_{ij}=a \mid \boldsymbol{x}\right]/P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}{P\left[y_{ij}=b \mid \boldsymbol{x}\right]/P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}\right) \\
&= \log\left(\frac{P\left[y_{ij}=a \mid \boldsymbol{x}\right]}{P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}\right) - \log\left(\frac{P\left[y_{ij}=b \mid \boldsymbol{x}\right]}{P\left[y_{ij}=1 \mid \boldsymbol{x}\right]}\right) \\
&= (\mu_a - \mu_b) + (\phi_a - \phi_b)\boldsymbol{\delta}'\boldsymbol{x}.
\end{aligned}$$

The relation between these two response categories is established by the scores parameters ϕ_a and ϕ_b. Thus, the larger the difference $(\phi_a - \phi_b)$ in absolute value, the more the odds of a and b are influenced by the predictor variables $\boldsymbol{x}$. In that manner, when the scores $\{\phi_k\}$ are constrained with the ordered increasing constraint (2.2), the effect of the covariates $\boldsymbol{x}$ is higher as the response categories increase. Additionally, the value of the response variable y_{ij} behaves as an ordinal response according to the value of $\boldsymbol{\delta}'\boldsymbol{x}$ when the scores $\{\phi_k\}$ are constrained. In other words, the larger the effect $\boldsymbol{\delta}$ of the covariates, the more the response variable y_{ij} has the propensity to be assigned to higher categories. Figure 2.1 shows an illustration of the $q = 4$ curves for the corresponding ordinal category probabilities $P[y_{ij} = k]$ from the boys with disturbed dreams example (see Table 1.1) and when $\delta_{\text{age}} > 0$. When $\delta_{\text{age}} < 0$, the labels in Figure 2.1 reverse order.

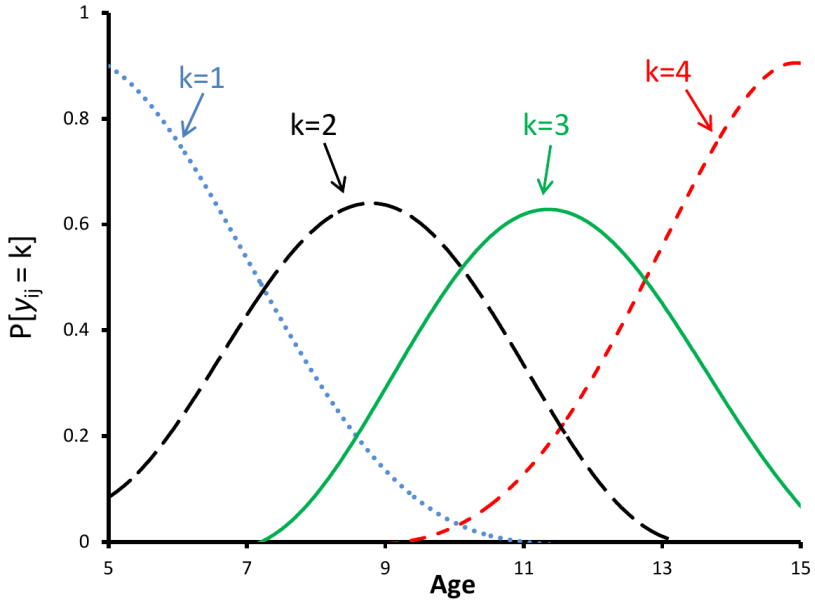

Figure 2.1: ***Interpretation:*** *Depiction of the category probabilities in the ordinal stereotype model from the boys with disturbed dreams example (see Table 1.1) with the ordinal score constraint* $0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1$ *and the age effect coefficient* $\delta_{\text{age}} > 0$*. Note that the four probabilities sum to 1 at any particular x-value (i.e. at any particular age).*

The stereotype model gives a way of estimating how close two adjacent categories are, e.g. k and $k+1$, based on how close their scores are, i.e. ϕ_k and ϕ_{k+1}. For instance, if the scores in the example described in the Table 1.1 are $\hat{\phi}_1 = 0$, $\hat{\phi}_1 = 0.32$, $\hat{\phi}_3 = 0.45$ and $\hat{\phi}_4 = 1$ means that the stereotype model implies that the adjacent categories "low severe" and "medium severe" are "close" given that their corresponding score parameters are close to each other. In the case that the scores between these two categories are the same, $\phi_a = \phi_b$, the corresponding logit for those two response categories is the constant $\mu_a - \mu_b$ and, therefore, the covariates $\boldsymbol{x}$ do not distinguish between them. In that case, we could collapse them in our data, as the categories have the same score. For instance, following the same example about the disturbed dreams of boys, the predictor variable "age" is not a useful covariate for predicting between "low" and "medium" degree of suffering in a boy if the scores are $\hat{\phi}_1 = 0$, $\hat{\phi}_4 = 1$ and $\hat{\phi}_2 = \hat{\phi}_3$ and, therefore, we could combine the categories "low severe" and "medium severe" into one single response category. In the same manner, one way to evaluate if two adjacent categories k and $k+1$ are distinguishable is by inspecting the standard

errors for their corresponding scores. Overlapping confidence intervals around the scores ϕ_k and ϕ_{k+1} may give evidence that ordinal categories k and $k+1$ are not distinguishable and we can collapse them into a single category.

An advantage of the stereotype model is that it requires a smaller number of parameters to be estimated than the baseline-category logit model (1.4) or the multinomial logistic regression model (1.5). As a way of illustrating this, we consider again the example of the severity of disturbed dreams in boys with $q = 4$ and with the supposition that there are three predictor variables as we did for the multinomial logistic regression model (see (1.6)). The pairs of log odds for the stereotype model are as follows,

$$\log\left(\frac{P\left[Y=2 \mid \boldsymbol{x}\right]}{P\left[Y=1 \mid \boldsymbol{x}\right]}\right) = \mu_2 + \phi_2(\delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3),$$

$$\log\left(\frac{P\left[Y=3 \mid \boldsymbol{x}\right]}{P\left[Y=1 \mid \boldsymbol{x}\right]}\right) = \mu_3 + \phi_3(\delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3), \text{ and} \tag{2.5}$$

$$\log\left(\frac{P\left[Y=4 \mid \boldsymbol{x}\right]}{P\left[Y=1 \mid \boldsymbol{x}\right]}\right) = \mu_4 + (\delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3),$$

which is more parsimonious (8 parameters) than the multinomial logistic regression model (12 parameters, see (1.6)). Moreover, as a result of the $\{\phi_k\}$ parameters, the ordered stereotype model is more flexible than the models including the proportional odds structure such as the version for the cumulative logit model (1.1). However, the parameters are more difficult to estimate due to the intrinsic nonlinearity which arises from the product of parameters $\phi_k \boldsymbol{\delta}' \boldsymbol{x}$ in the predictor. We discuss this in the next section.

2.2 Fitting the Ordered Stereotype Model

Despite its parsimonious structure, the stereotype model for ordered responses (eq. (2.1) and (2.2)) has rarely been used in applied research (see Kuss (2006) for some exceptions) compared to other equivalent models such as the proportional odds model version of the cumulative logit, the adjacent-categories logit model and the multinomial logistic regression (eq. (1.1), (1.3) and (1.5) respectively).

This lack of use may have arisen from the lack of standard software for model fitting, the intrinsic non-linear structure of the predictor, and the requirement of multiple constraints in the parameter space to ensure the identifiability in the parameters.

There have however been considerable recent developments of macros and functions in standard software to estimate the stereotype model. Lunt (2001) developed a STATA module called SOREG that implements this model, Greenland (1994) stated the potential of using reduced-rank multinomial logistic models (RR-MLM) as a special case of the stereotype model without including the ordinal constraint, and Yee and Hastie (2003) fitted RR-MLM by using the VGAM (Vector Generalized Additive Model) package for **R** (Yee (2008)). Kuss (2006) modified several standard procedures in SAS to obtain the maximum likelihood estimator (MLE) values by applying direct maximisation of the likelihood function. A ML estimation procedure, whether by using the EM algorithm or some numerical optimisation technique, requires a good set of parameter starting points to initialize the algorithm optimiser and obtain reliable MLE values. A possible procedure to specify good starting points may be to fit the multinomial logistic regression model (1.5) first and use the estimated parameters for the baseline category (e.g. the first category) as starting points. In that manner, $\boldsymbol{\delta_1}$ and $\{\mu_2, \ldots \mu_q\}$ from the multinomial logistic regression model would be the starting points of $\boldsymbol{\delta}$ and $\{\mu_k\}$ respectively for the stereotype model. Similarly, the starting points for the score parameters $\{\phi_k\}$ may be obtained from the relationship between both models, given by $\boldsymbol{\delta_k} = \phi_k \boldsymbol{\delta}$ for $k = 2, \ldots, q$. However, these computational adjustments in SAS are based on unconstrained optimisation methods and therefore they ignore the ordinal increasing constraint $\phi_1 < \ldots < \phi_q$ in the score parameters $\{\phi_k\}$.

The stereotype model cannot be considered as a generalized linear model (GLM) due to the non-linear combination of parameters (i.e. a multiplicative combination of $\{\phi_k\}$ and $\{\delta_p\}$: $y_{ij} = \mu_k + \phi_k \sum_{\ell=1}^{p} \delta_\ell x_{\ell ij}$). Therefore, its inference is infeasible using software for GLMs. There are several suggestions for fitting this model in the literature. In his seminal paper on the stereotype model, Anderson (1984) recommended a model fitting procedure consisting of direct iterative optimisation of the likelihood function but the paper does not include an explicit procedure or code. Holtbrugge and Schumacher (1991) proposed a method to estimate the parameters in the stereotype model by using an iteratively reweighted least square algorithm and Feldmann and König (2002) proposed a maximum li-

kelihood parameter estimation based on discriminant analysis. Greenland (1994) proposed an alternating algorithm based on two iterative steps. At the first step, the score parameters $\{\phi_k\}$ are kept fixed and then the parameters $\boldsymbol{\delta}$ in the predictor can be estimated. As was mentioned above, this treats the stereotype model as a RR-MLM and then it is possible to fit it by using standard generalized linear model software via constrained polytomous logistic regression. At a second step, $\widehat{\boldsymbol{\delta}}'\boldsymbol{x}$ is treated as an estimated scalar predictor and it is possible to estimate $\{\phi_k\}$ conditional on $\boldsymbol{\delta}' = \widehat{\boldsymbol{\delta}}'$ by the same model fitting procedure. This iterative procedure has the drawbacks that the ordinal increasing constraint is not included in the fitting procedure, the convergence at the true MLE values is not guaranteed, and the standard-error estimates obtained by the standard software at convergence are not valid. Instead, the author recommends the use of a Monte Carlo simulation from the estimated model for computing confidence intervals and p-values.

As was mentioned above, another difficulty in the implementation of the ordered stereotype model is the imposition of the monotone increasing constraint (2.2) on the score parameters $\{\phi_k\}$ in the fitting procedure. Given that difficulty, Greenland (1994) and Lunt (2005) suggested several alternatives with the purpose of avoiding the estimation of $\{\phi_k\}$ as for example determining $\{\phi_k\}$ in advance based on background information such as previous and pilot studies, so that then we can use generalized linear models. As an alternative to the classical frequentist approach, Ahn et al. (2009) presented comprehensive Bayesian inference and model comparison procedure for fitting the ordinal stereotype model applied including the constraint on $\{\phi_k\}$ in the case-control studies. In addition, Ahn et al. (2011) showed two methods for parameter estimation in this model in the presence of missing exposure data by using a Monte Carlo approach as well as an expectation/conditional maximisation algorithm. More recently, the fitting of the ordered stereotype model based on the iterative alternating algorithm of Greenland (1994) is proposed in Preedalikit (2012).

2.3 Ordered Stereotype Model Including Clustering

The structure of the predictor in the ordered stereotype model (2.1) can include the predictor variables $\boldsymbol{x}$ as numerical covariates, or alternatively they may sim-

ply depend on the row and/or column of the observation y_{ij}. We consider this latter situation and build up $\boldsymbol{\delta'x}$ only taking into account the row and column effects by using a linear formulation. To do this, we define $\{\alpha_1, \ldots, \alpha_n\}$ and $\{\beta_1, \ldots, \beta_m\}$ as the sets of parameters quantifying the main effects of the n rows and m columns respectively, and the set $\{\gamma_{11}, \ldots, \gamma_{nm}\}$ are the interaction effects of the different rows and columns. In this way, we can formulate the following saturated model

$$\begin{aligned} &\log\left(\frac{P\left[y_{ij}=k\right]}{P\left[y_{ij}=1\right]}\right) = \mu_k + \phi_k(\alpha_i + \beta_j + \gamma_{ij}), \\ &k = 2, \ldots, q, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \end{aligned} \tag{2.6}$$

where $\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{m} \beta_j = 0$ and we impose sum-to-zero constraints on each row and column of the association (or pattern detection) matrix γ. This model has $2q + nm - 4$ independent parameters, i.e. more parameters than the nm observations in the matrix Y, when $q > 2$. The relationship between models (2.3) and (2.6) is shown in Appendix A.1. Since this model is overparametrised, the most common submodels to formulate from the saturated model are the main effect model ($\gamma_{ij} = 0$, with $2q + n + m - 5$ parameters), the row effect model ($\beta_j = \gamma_{ij} = 0$, $2q + n - 4$ parameters), the column effect model ($\alpha_i = \gamma_{ij} = 0$, $2q + m - 4$ parameters) and the null model ($\alpha_i = \beta_j = \gamma_{ij} = 0$, $2q - 3$ parameters).

The main problem with the model in (2.6) is of course that the specific row and column effects in this suite of models over-parameterizes the data structure. This model is not parsimonious and it requires a lot of parameters for describing all the effects. A way to reduce the dimensionality of the problem is to introduce fuzzy clustering via finite mixtures. Hence, we obtain the following model formulation including row clustering, column clustering or biclustering.

- Row clustering

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

- Column clustering

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

- Biclustering

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

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

Finally, in the same way as before, we can formulate the probability of the

data response y_{ij} being equal to the category k conditional on the appropriate clustering as,

- Row clustering

$$\begin{aligned}\theta_{r_ijk} &= P\left[y_{ij} = k \mid i \in r\right] = \frac{\exp(\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj}))}{\sum_{\ell=1}^{q} \exp(\mu_\ell + \phi_\ell(\alpha_r + \beta_j + \gamma_{rj}))}, \\ k &= 1, \ldots, q, \quad r = 1, \ldots, R, \quad j = 1, \ldots, m.\end{aligned} \tag{2.7}$$

- Column clustering

$$\begin{aligned}\theta_{ic_jk} &= P\left[y_{ij} = k \mid j \in c\right] = \frac{\exp(\mu_k + \phi_k(\alpha_i + \beta_c + \gamma_{ic}))}{\sum_{\ell=1}^{q} \exp(\mu_\ell + \phi_\ell(\alpha_i + \beta_c + \gamma_{ic}))}, \\ k &= 1, \ldots, q, \quad c = 1, \ldots, C, \quad i = 1, \ldots, n.\end{aligned} \tag{2.8}$$

- Biclustering

$$\begin{aligned}\theta_{r_ic_jk} &= P\left[y_{ij} = k \mid i \in r, j \in c\right] = \frac{\exp(\mu_k + \phi_k(\alpha_r + \beta_c + \gamma_{rc}))}{\sum_{\ell=1}^{q} \exp(\mu_\ell + \phi_\ell(\alpha_r + \beta_c + \gamma_{rc}))}, \\ k &= 1, \ldots, q, \quad r = 1, \ldots, R, \quad c = 1, \ldots, C.\end{aligned} \tag{2.9}$$

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

2.4 Basic Models. Likelihoods

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

2.4.1 Row Clustering

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

rows into R clusters:

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

where Ω is the parameter vector for the case of row clustering, π_{r_i} is the *a priori* row membership probability of row i and $\theta_{r_i jk}$ is the probability of the data response defined in (2.7). Assuming independence among rows and, conditional on the rows, independence over the columns, we can simplify the previous incomplete data likelihood to

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

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

$$Z_{ir} = I(i \in r) = \begin{cases} 1 & \text{if } i \in r \\ 0 & \text{if } i \notin r \end{cases} \qquad i = 1, \ldots, n, \quad r = 1, \ldots, R, \tag{2.10}$$

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

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

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

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

These indicator latent variables fulfill the following convenient identity

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

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

$$l_c(\Omega \mid \{y_{ij}\}, \{z_{ir}\}) = \sum_{i=1}^{n}\sum_{r=1}^{R} z_{ir}\log(\pi_r) + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{r=1}^{R} z_{ir}I(y_{ij}=k)\log\left(\theta_{r_ijk}\right). \tag{2.11}$$

2.4.2 Column Clustering

The model for the case of clustering the columns but not the rows is similar. It assumes independence among columns and, conditional on the columns, independence over the rows. Analogous to Z_{ir} for row clustering (see (2.10)) we define the following indicator latent variables for the unknown data

$$X_{jc} = I(j \in c) = \begin{cases} 1 & \text{if } j \in c \\ 0 & \text{if } j \notin c \end{cases} \qquad j = 1, \ldots, m, \quad c = 1, \ldots, C. \tag{2.12}$$

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

$$l_c(\Omega \mid \{y_{ij}\}, \{x_{jc}\}) = \sum_{j=1}^{m}\sum_{c=1}^{C} x_{jc}\log(\kappa_c) + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{c=1}^{C} x_{jc}I(y_{ij}=k)\log\left(\theta_{ic_jk}\right), \tag{2.13}$$

where Ω is the parameter vector for the case of column clustering, κ_c is the *a priori* column membership probability and θ_{ric_jk} is the probability of the data response defined in (2.8).

2.4.3 Biclustering

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

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

Here Ω is the parameter vector for the case of biclustering and $\theta_{r_i c_j k}$ is the probability of the data response expressed in (2.9). Assuming independence among rows and, conditional on the rows, independence over the columns, we can simplify the previous incomplete data likelihood to

$$L(\Omega \mid \{y_{ij}\}) = \sum_{c_1=1}^{C} \cdots \sum_{c_m=1}^{C} \kappa_{c_1} \cdots \kappa_{c_m} \prod_{i=1}^{n} \left[\sum_{r=1}^{R} \pi_r \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\theta_{r_i j k}\right)^{I(y_{ij}=k)} \right], \tag{2.14}$$

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

$$L(\Omega \mid \{y_{ij}\}) = \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \cdots \pi_{r_n} \prod_{j=1}^{m} \left[\sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \left(\theta_{i c_j k}\right)^{I(y_{ij}=k)} \right]. \tag{2.15}$$

We define the unknown data through the indicator latent variables described in (2.10) and (2.12). Consequently, the complete data log-likelihood of this model using the known data $\{y_{ij}\}$ and the unknown data $\{z_{ir}\}$ and $\{x_{jc}\}$ is as follows:

$$\begin{aligned} l_c(\Omega \mid \{y_{ij}\}, \{z_{ir}\}, \{x_{jc}\}) = & \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{r=1}^{R} \sum_{c=1}^{C} z_{ir} x_{jc} I(y_{ij} = k) \log\left(\theta_{r_i c_j k}\right) \\ & + \sum_{i=1}^{n} \sum_{r=1}^{R} z_{ir} \log\left(\pi_r\right) + \sum_{j=1}^{m} \sum_{c=1}^{C} x_{jc} \log\left(\kappa_c\right). \end{aligned} \tag{2.16}$$

We estimate the MLEs from this expression by using the EM algorithm. In the E-step, the expected value of the first term is approximated using the variational approximation employed by Govaert and Nadif (2005) (see Appendix A.3 for details). With the aim of ensuring a solution avoiding approximations, we use the resulting MLEs from the EM algorithm as starting points to numerically maximise the incomplete-data likelihood (2.14) (or (2.15)). We note that during the maximisation a convenient transformation for the row and column membership parameters $\{\pi_r\}$ and $\{\kappa_c\}$ is $s_r = \text{logit}(\pi_r / \sum_{\ell=r}^{R} \pi_\ell)$ for $r = 1, \ldots, R-1$ and $q_c = \text{logit}(\kappa_c / \sum_{\ell=c}^{C} \kappa_\ell)$ for $c = 1, \ldots, C-1$ respectively. This transformation means that the parameters s_r and q_c are unconstrained, taking values over the whole real line.

2.5 Estimation of the Parameters

As we introduced in Section 1.2.1, a powerful and common method for finding the maximum likelihood solution for model with missing information and, therefore, involving *latent variables* is called the *expectation-maximisation* algorithm or EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997). One of the most common uses of the EM algorithm is in the case of the estimation of the parameters for a finite mixture-density model with incomplete data which in this case is the actual unknown cluster membership of each row and/or column. Therefore, we can derive estimates of the parameters of the model expressed in (2.7), (2.8) and (2.9) by using the EM algorithm, taking into account the actual unknown cluster membership of each row and/or column. This method performs a fuzzy assignment of rows and/or columns to clusters based on the posterior probabilities. In this section, we develop this in detail for the case of clustering the rows but not the columns. It has an easy interpretation which helps explain our methodology. The development for other two cases: clustering the columns but not the rows and biclustering are described in the Appendices A.2 and A.3.

2.5.1 The Expectation Step (E-Step). Row Clustering

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

$$\begin{aligned} Q(\Omega \mid \Omega^{(t-1)}) &= E_{\{z_{ir}\}|\{y_{ij}\},\Omega^{(t-1)}}\left[\ell_c(\Omega \mid \{y_{ij}\},\{z_{ir}\})\right] \\ &= \sum_{i=1}^{n}\sum_{r=1}^{R} \log(\pi_r^{(t-1)}) E\left[z_{ir} \mid \{y_{ij}\}, \Omega^{(t-1)}\right] \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{r=1}^{R} I(y_{ij}=k) \log\left(\theta_{rjk}^{(t-1)}\right) E\left[z_{ir} \mid \{y_{ij}\}, \Omega^{(t-1)}\right]. \end{aligned} \tag{2.17}$$

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

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

and applying Bayes' rule to this expression we obtain

$$\begin{aligned}\widehat{Z}_{ir}^{(t)} = P\left[z_{ir} = 1 \mid \{y_{ij}\}, \Omega^{(t-1)}\right] &= \frac{P\left(\{y_{ij}\}, \Omega^{(t-1)} \mid z_{ir} = 1\right) P\left(z_{ir} = 1\right)}{\sum_{\ell=1}^{R} P\left(\{y_{ij}\}, \Omega^{(t-1)} \mid z_{i\ell} = 1\right) P\left(z_{i\ell} = 1\right)} \\ &= \frac{\widehat{\pi}_r^{(t-1)} \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\widehat{\theta}_{rjk}^{(t-1)}\right)^{I(y_{ij}=k)}}{\sum_{\ell=1}^{R} \left\{\widehat{\pi}_\ell^{(t-1)} \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\widehat{\theta}_{\ell jk}^{(t-1)}\right)^{I(y_{ij}=k)}\right\}}.\end{aligned} \tag{2.18}$$

This is the expected value of the latent variable Z_{ir} which defines the posterior probability that row i is in group r once we have observed $\{y_{ij}\}$. Finally, we complete the E-step by substituting the previous expression in the complete data log-likelihood at the iteration t expressed in (2.17),

$$\widehat{Q}(\Omega \mid \Omega^{(t-1)}) = \sum_{i=1}^{n} \sum_{r=1}^{R} \widehat{Z}_{ir}^{(t)} \log(\widehat{\pi}_r^{(t-1)}) + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{r=1}^{R} \widehat{Z}_{ir}^{(t)} I(y_{ij} = k) \log\left(\widehat{\theta}_{rjk}^{(t-1)}\right). \tag{2.19}$$

2.5.2 The Maximisation Step (M-step). Row Clustering

The M-step of the EM algorithm is the global maximisation of the previous expression (2.19) obtained in the E-step. For the case of finite mixture models, the updated estimation of the term containing the row-cluster proportions $\{\pi_1, \ldots \pi_R\}$ and the one containing the rest of the parameters Ω are computed independently. Thus, the M-step has two separate parts.

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

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

However, the data z_{ir} are unobserved in our case. Because of this, we use their conditional expectation which we found in the E-step (2.18) to replace in the pre-

vious expression for the iteration t,

$$\widehat{\pi}_r^{(t)} = \frac{1}{n}\sum_{i=1}^{n} E\left[z_{ir} \mid \{y_{ij}\}, \Omega^{(t-1)}\right] = \frac{1}{n}\sum_{i=1}^{n} \widehat{Z}_{ir}^{(t)}, \qquad r = 1, \ldots, R. \tag{2.20}$$

Secondly, to estimate the remaining parameters Ω, we must numerically maximise the conditional expectation of the complete data log-likelihood (2.17). In the case of row clustering,

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

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

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

Other criteria might be applied (see e.g. Abbi et al. (2008) for a review of various stopping criteria). A disadvantage of mixture modelling is that the associated likelihood surface may be multimodal. A comprehensive search over different starting points is used to avoid finding only a local maximum. Particularly in our case, the iterative process is repeated 10 times with random starting points and the best MLE (those that lead to higher log-likelihood value) are kept. We have run experiments testing up to 100 random starting points and it was sufficient with 10 repetitions to avoid convergence to local optima. The random starting points can be generated with simulated annealing (Kirkpatrick et al., 1983; Zhou and Lange, 2010), a good option to avoid being locked into a local maximum.

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

2.5.3 Reparametrisation of the Score Parameters

The increasing constraint that enforces the scores $\phi_1, \ldots, \phi_q$ to be increasing in the stereotype model defined in (2.2) must be imposed during the estimation procedure. Such a constraint is complex to impose during optimisation and hence for convenience we reparametrise $\phi_1, \ldots, \phi_q$ as follows.

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

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

We then set

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

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

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

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

2.6 Discussion

The ordinal stereotype model and its formulation including fuzzy clustering via finite mixtures has been introduced in this chapter. Additionally, model fitting for the clustering version by using the EM algorithm was presented. Two common drawbacks of finite mixture-density likelihood are the existence of multiple

maxima and the occurrence of singularities. As a consequence of the multiple convergence, the EM algorithm may find local maximum likelihood estimates instead of the global maximum estimates (McLachlan and Krishnan, 1997). In that manner, a solution is running the EM algorithm repeatedly using different and widespread sets of initial values. The observation of the same maximum likelihood estimates from diverse initial points increases confidence that the solution is a global maximum. On the other hand, the appearance of singularities may be common when the number of parameters to estimate is large in relation to the sample size. It produces degenerate distributions where the likelihood function becomes infinite. Possible solutions to this problem would be to constrain the range of the row (column) effect values or use a Bayesian estimation method as alternative to the EM algorithm (Fraley and Raftery, 2007). The description of a Bayesian estimation method for our approach is described in Chapters 7 and 8.

Chapter 3

Model Selection for Ordinal Finite Mixtures

3.1 Introduction

One of the key questions in the use of mixture models to provide a model-based clustering is concerning the choice of the number of mixture components most suitable to the data set. Since each component in the mixture corresponds to a different cluster in the data, the conclusions obtained from the estimated mixture model may be inaccurate when the estimated number of clusters is incorrect.

There have been several approaches via finite mixture models to solve the classification problem of deciding how many clusters there are in the data. However, the majority of them are based on clustering continuous outcome variables and, in particular, on the multivariate normal distribution (see McLachlan (1982); McLachlan and Basford (1988); Fraley and Raftery (2002) for an extensive review). Therefore, there has so far been minimal research on model selection for finite mixture models in the ordinal case and, in particular, when mixture components are based on ordinal stereotype distributions. An example of research on ordinal variables for the mixture components is the work by Lanning and Bozdogan (2004, Chapter 21) which is focused on the proportional odds and the nested cumulative link model.

One of the most well documented approaches to the model choice problem is based on assessing the number of modes in the data (see an extensive review by McLachlan and Peel (2000, Chapter 6)). A criticism of this approach that the

mixture components must be well separated in order to be distinguished. Therefore, the finite mixture model once it is fitted might underestimate the number of groups in the data. The specification of a parametric distribution family for the mixture components may overcome this drawback. However, it is important to take into account that the estimation of the number of mixture components might not reflect the actual number of groups in the data. For instance, if the data has a skewed distribution within some of the groups, then there may not be an one-to-one correspondence between groups and mixture components. This issue has been considered by McLachlan and Peel (2000, Chapter 6) in normal mixture models. In addition, Ray and Ren (2012) is a useful reference on the misleading results that might be obtained using mode counting to assess the number of components.

Other standard approaches to the problem of determining the number of clusters without prior knowledge of their composition have been considered in the literature (see e.g. Engelman and Hartigan (1969); Bock (1974); Bozdogan (1993); Fraley and Raftery (1998)). A common methodology consists of using the EM algorithm to estimate the composition of the finite mixture model for various assumed numbers of clusters. Once the set of candidate models is estimated, the next step is to use an information criterion to select the best number of mixture components. The use of information criteria provides an objective method for the selection of a best approximating model for the data while allowing a direct ordering of the candidate models for comparison. Furthermore, unlike significance testing, this allows comparison of more than two models at the same time. A hard clustering structure is then created by assigning each observation from the data to the cluster to which it is most likely to belong a posteriori, conditionally on the selected model and its estimated parameters. This is the so-called mixture ML approach which is equivalent to the more computationally complex classification ML approach (see McLachlan (1982) and McLachlan and Peel (2000, Section 2.2.1)) with the additional assumption that the probabilities of component memberships is an unobservable random sample as described in Section 1.2.1. Examples showing that this methodology can perform better than standard approaches are given in Fraley and Raftery (1998) for the case of multivariate normal mixture components. In this chapter, we use this methodology for selecting the best model or set of models in our proposed likelihood-based approach based on ordinal stereotype fuzzy clustering via a finite mixtures model.

Although there is research concerning which current available information criteria are most suitable in order to select the number of clusters for finite mixture models, the performance of information criteria where the mixture components are based on ordinal stereotype distributions is a question that still remains. For this reason, we set up simulation experiments with the aim of comparing the performance of several information criterion measures and choosing the most suitable (or set of them) for our approach. The knowledge of the best criteria in advance will give valuable guidance to practitioners who might later apply our methodology. There are similar empirical comparisons in the literature. See for example works from Fonseca and Cardoso (2007) and McLachlan and Ng (2000).

The chapter is organized as follows: in Section 3.2, we review the information criteria this chapter is focused on. A simulation study comparing their performance is presented in Section 3.3. In Section 3.4, the best information criteria from the experimental study are tested by applying our approach to a real-life dataset which is known to be composed of a certain number of clusters. Finally, conclusions and future research are described in Section 3.5.

3.2 Model Selection Criteria in Clustering

Model selection is an important stage in any data analysis. As the full truth might not be able to be represented in a model, the selection of an estimated best approximating model from a set of candidate models is critical for statistical inference. In a clustering context, there are two main approaches to the comparison of a set of candidate likelihood-based models once they are fitted, in order to decide which one (or group of them) best approximates the (unknown) true model. One approach is to carry out a hypothesis test by using the likelihood ratio test as a test statistic (LRT). Another approach uses information criteria which are based on a penalised form of the likelihood function where the penalty increases as the number of parameters increases. A review of these methods is given in the following two sections.

3.2.1 Likelihood Ratio Tests

A simple hypothesis test regarding the number of groups in a mixture model is formulated as:

$$H_0 : g = g_0 \text{ vs. } H_\text{a} : g = g_\text{a},$$

where g is the number of mixture components and g_0 and g_a are two possible number of group values. A common statistic to select the best model in a set of candidate models is the likelihood radio test (LRT) which has test statistic

$$\text{LLR} = \log\left(\frac{\text{The maximised likelihood with } g = g_0}{\text{The maximised likelihood with } g = g_\text{a}}\right), \tag{3.1}$$

which has known asymptotic properties under certain regularity conditions (Wilks, 1938). However, a drawback of the use of this test for mixture densities is that it does not lead to a suitable significance test. This occurs because the null hypothesis under test is defined on the boundary of the parameter space and, consequently, the required regularity conditions do not hold for -2LLR to have its usual asymptotic chi-square distribution. In particular in finite mixtures, the mixing proportions of two mixture components become unidentifiable when the components coincide. The consequence of this is that the LRT tends to overestimate the number of clusters (Stahl and Sallis, 2012; Everitt et al., 2011). Self and Liang (1987) derived approximate asymptotic null distributions for modified LLRs which are valid at the boundary of the parameter space. For example, if testing between g_0 and g_a components where $g_\text{a} = g_0 + 1$, the null distribution is approximately 50:50 mixture of zeros and χ^2_1 values. However their results, while computationally easy to implement, refer to normal distributions, and may not be applicable to ordinal data.

There has been a lot of published research formulating results on the null distribution of the LLR for the finite mixture model through simulation and bootstrapping studies (see the review in McLachlan and Peel (2000, Section 6.5) and Everitt et al. (2011, Section 6.5.1)). One of the most common ways may be using randomisation tests to obtain the asymptotic null distribution (McLachlan, 1987; Manly, 2007; Gotelli and Graves, 1996). However once again, most of the results reported have been focused on mixtures whose components are densities from continuous variables. In addition, a common drawback is that the use of LRT might be computationally demanding because it requires bootstrapping

to obtain the p-value. Therefore, there is a lack of research on this area focused on mixtures based on densities from ordinal variables and it might be a field to explore for future research.

In the next section, we introduce an alternative way to select the number of mixture components based on information theoretical methods.

3.2.2 Description of Information Criteria

The use of model selection procedures based on information criteria methods started when Akaike (1973) introduced a relationship between the Kullback-Leibler distance [1] ($\mathrm{D_{KL}}$, Kullback and Leibler (1951)) and the log-likelihood function. The expression of the $\mathrm{D_{KL}}$ between two models f (true model) and g (approximating model) in the case of continuous distributions is

$$\begin{aligned}\mathrm{D_{KL}} &= \int f(Y) \log\left(\frac{f(Y)}{g(Y \mid \Omega)}\right) \mathrm{d}Y \\ &= \int f(Y) \log\left(f(Y)\right) \mathrm{d}Y - \int f(Y) \log\left(g(Y \mid \Omega)\right) \mathrm{d}Y,\end{aligned}$$

where Y denotes the data being modeled and Ω denotes the parameters in the approximating model g. Note, each of the two terms on the right is a statistical expectation with respect to the true model f. Thus, the $\mathrm{D_{KL}}$ distance can be expressed as

$$\begin{aligned}\mathrm{D_{KL}} &= E_f\left[\log(f(Y))\right] - E_f\left[\log\left(g(Y \mid \Omega)\right)\right] \\ \mathrm{D_{KL}} - E_f\left[\log(f(Y))\right] &= -E_f\left[\log\left(g(Y \mid \Omega)\right)\right].\end{aligned} \tag{3.2}$$

The first expectation $E_f\left[\log(f(Y))\right]$ depends only on the unknown true model f. Moreover, it has an unknown value but is constant for any approximating model g. Thus, the second expectation $E_f\left[\log\left(g(Y \mid \Omega)\right)\right]$ becomes the quantity of interest. Therefore, the quantity $\mathrm{D_{KL}} - E_f\left[\log(f(Y))\right]$ is a measure of the loss of information when a particular candidate model g is used to approximate the unknown true model f, if the estimation of $E_f\left[\log\left(g(Y \mid \Omega)\right)\right]$ is possible. In this manner, all candidate models are scored regarding their relative information loss and the best model is that with lowest loss. Unlike the LRT, information criteria

[1] Kullback-Leibler "distance" is not a proper distance metric, as it is not symmetric. Some authors recently refer to it as the Kullback-Leibler discrepancy.

allow quantification of the differences among a set of candidate models and there may be not a single best model.

Several information criteria have been developed by measuring the loss of information as a balanced penalty described by the lack of fit (based on the maximised log-likelihood function) plus a lack of parsimony (using measures of model complexity). The general formula of an information criterion CRI is as follows:

$$\text{CRI} = -2\ell + P \tag{3.3}$$

where ℓ is the maximised log-likelihood and P corresponds to the penalty term, and a lower value of CRI indicates a better model.

The first term of this equation decreases when the model complexity increases, improving the fit, and the penalty term P increases when the model complexity (e.g. number of parameters) increases. Therefore, the first term rewards goodness of fit whereas the penalty term P discourages over-fitting in the model estimation. As increasing the number of parameters in the model always improves even minimally the goodness of the fit, this formulation of information criteria implements a trade-off between the fitted description of the data and the number of parameters of the model. Thus, an information criterion selects the best model in a finite set of candidate models. However it does so even when all those models are very poor, i.e. CRI is a relative not an absolute measure. Therefore, it is important to include well founded models in the candidate set, and to include checks for model assessments adequately.

Some of the most common information criteria for the estimation of mixture models are described below, and their definitions are collected in Table 3.1 on page 48.

Akaike's information criterion (AIC)

The most commonly used information criterion is Akaike's information criterion (AIC). It was introduced in Akaike (1973) and is founded on information theory. The paper proposes the use of an estimate of the Kullback-Leibler information as a fundamental basis for model selection. Aikake found a rigorous way to estimate Kullback-Leibler distance by means of the maximum point of the empirical log-likelihood function (Burnham and Anderson, 2002). In that manner, AIC provides an estimation of the relative distance between the fitted model and the

unknown true model.

AIC is formulated from equation (3.3) with a penalty term of $P = 2K$ where K is the number of free parameters. Thus,

$$\text{AIC} = -2\ell + 2K.$$

There is a relationship between AIC and LRT when the analysis regards two nested models i and j with model i a restriction of the fuller model j (see e.g. Haefner (2005, Section 8.4)):

$$\text{LLR} = \text{AIC}_i - \text{AIC}_j - 2(K_i - K_j) = 2(\ell_j - \ell_i),$$

where K_i and K_j are the number of free parameters, and ℓ_i and ℓ_j are the maximised log-likelihood in each model.

The choice of the number of mixture components when finite mixture models are in the set of candidate models is the primary interest. The use of AIC as a model selection criterion causes concern because the maximised log-likelihood ℓ lies on the boundary of the parameter space and, consequently, the regularity conditions fail. Some authors observed that AIC tends to overestimate the correct numbers of components for mainly multivariate normal components in the mixture context (Soromenho, 1994; Celeux and Soromenho, 1996). Thus, it is natural to think that AIC should be modified in that case. Despite this, Burnham and Anderson (2002, Section 6.9.6) advocate the use of AIC without alterations for mixture models. However, AIC must be used carefully because the model selection procedure entails some traps that must be avoided. The general idea is to use AIC by choosing the correct number of free parameters. For example, a 4-component mixture model might collapse to a 3-component mixture model when one of the mixing proportions lies on a parameter boundary (e.g. $\pi_3 = 0$). However, a 3-component mixture model might have already been included within the set of candidate models. Thus, this set would be redundant as it would include two different models with 3-component mixtures. Therefore, the set of candidate models must be adjusted. One criterion of adjustment can be to select the model with higher likelihood, among the redundant models. We have used this principle in the simulation study which is described in Section 3.3.

Modified AIC criterion (AIC3)

The AIC3 criterion was introduced by Bozdogan (1994) as a modification of AIC by increasing the penalty term regarding the number of free parameters in the model. Thus,

$$\text{AIC3} = -2\ell + 3K.$$

This measure has been found to perform empirically better in the context of multivariate normal and Bernoulli mixture models (Andrews and Currim, 2003) where the fitted parameters may lie on the parameter space boundary as in the case of finite mixtures.

Fonseca and Cardoso (2007) set up a comprehensive simulation study in the case of categorical variables and when a finite mixture model of conditionally independent multinomial distributions is adopted. As a result of their experiments, AIC3 had best performance than AIC in terms of the detection of the information criterion to discover the true number of clusters.

Corrected AIC criterion when sample size is small (AIC_c)

AIC may perform poorly with small sample sizes and, particularly, when the number of parameters to estimate is large in relation to the sample size. The AIC_c criteria was proposed by Hurvich and Tsai (1989) based on Akaike (1973) paper as a correction of AIC when the sample size is small. To achieve this, AIC_c adjusts the penalty term in formula (3.3) including a greater penalty for extra parameters, $P = \frac{2K(K+1)}{n-K-1}$ where K is the number of free parameters and n the sample size.

Corrected AIC_c criterion when sample size increases (AIC_u)

McQuarrie et al. (1997) remarked that despite the small sample adjustment AIC_c still tends to overestimate the number of parameters as the sample size increases. They therefore proposed the criterion AIC_u which includes a larger penalty term $P = n \log\left(\frac{n}{n-K-1}\right)$. They showed that this new criterion is an approximate unbiased estimator of Kullback-Leibler distance and improves the results of AIC_c for moderate to large sample sizes.

Bayesian information criterion (BIC)

The criteria that have so far been described in this chapter are based on the assumption that there is an unknown true model and the aim is to select its best approximated candidate model. There is another point of view which is based on the assumption that not only does an exactly true model exist, but that it is included within the set of candidate models under consideration. Additionally, another implicit strong assumption from this context is that the true model is of fairly low dimension (i.e. K is less than 5 parameters). Several criteria have been developed based on this perspective (see Bozdogan (1987) for a complete review). The best-known is the Bayesian information criterion (BIC) and was proposed by Schwarz (1978). BIC is formulated from equation (3.3) with a penalty term of $P = K\log(n)$ where K is the number of free parameters and n the sample size. This modification of the penalty term arises from considering an equal prior probability on each candidate model and obtaining the asymptotic behavior of Bayes estimators. BIC penalizes complex models more heavily than AIC when $n > 8$, because then $\log(n) > 2$. BIC is not an estimator of Kullback-Leibler distance.

Support for BIC is mixed. For example, Fraley and Raftery (1998) note that there is considerable support for use of BIC for finite mixture models, Leroux (1992) shows the integrated classification criterion (ICL, Biernacki et al. (1998)) is an approximation to BIC which does not underestimate the true number of components, and Keribin (2000) has shown that BIC performs consistently in choosing the true number of components using a maximum penalised likelihood method under an appropriate penalisation sequence. However, most of those conclusions are based on the continuous case (especially normal mixture models) or asymptotic results.

There have been a number of criticisms of BIC. For example it has been found that BIC does not have good performance when the true model has a complex structure (Burnham and Anderson, 2002, Section 6.3.2.) and it tends to select models that are too simple in realistic situations. Umbach and Wilcox (1998) conducted a Monte Carlo simulation study to test how the AIC and BIC perform over different sample sizes. For sample sizes up to 100000, they found AIC performed better than BIC in terms of selecting the true model. The two criteria were tied at sample size 125000, and BIC had the best performance at sample sizes larger than 125000. This empirical study seems to suggest than BIC tends to select overly

simple models when the sample size is not large enough.

There have also been a number of comparisons of all three of BIC, AIC, and AIC_c (see e.g. Hjorth (1994, Section 3.7)). They note that BIC performs well asymptotically (i.e. increasing the sample size). However, the sample size would have to be very large in order to achieve a reasonable accuracy when more than one model is close to the true model.

Integrated classification likelihood criterion as approximation to BIC (ICL-BIC)

The integrated classification criterion (ICL) is a complete likelihood-based information criterion (also known as a classification-based information criterion) developed by Biernacki et al. (1998). McLachlan and Peel (2000) used ICL-BIC to refer to an approximated form of ICL which has the same form as BIC apart from the addition of the entropy penalisation term and showed that this criterion selected the true number of clusters in all 3 simulation normal continuous data sets that they considered. The ICL-BIC is expressed as:

$$\text{ICL-BIC} = -2\ell_c + K\log(n), \tag{3.4}$$

where ℓ_c is the maximised complete data log-likelihood function, K is the number of free parameters and n the sample size of the incomplete data.

Biernacki et al. (1998) argued that using the BIC for assessing mixture models to provide a model-based clustering presents some drawbacks. From a theoretical point of view, if the true model has $S' < S$ mixture components where S is the number of mixture components in the candidate model under consideration, then $S - S'$ of the mixing proportions will tend to zero as the sample size tends to infinity. Thus, the regularity conditions will fail because the mixing proportion estimates will lie on the boundaries of parameter space. Additionally, Biernacki and Govaert (1997) stated that if the true model is not in the family of models under consideration, BIC tends to overestimate the correct number of components regardless of the cluster separation. Biernacki et al. (1998) showed through numerical experiments that ICL-BIC increases the ability of the mixture model to give evidence for a clustering structure of the data.

Another equivalent way to formulate the ICL-BIC is taking $P = K\log(n) + 2\text{EN}(S)$ in equation (3.3), where $\text{EN}(S)$ is the entropy function (see Celeux and Soromenho (1996)) and S is the number of clusters. This function is a measure of

the ability of a particular S-component mixture model to allocate the data to the specified clusters. $\mathrm{EN}(S)$ measures the overlap of the mixture components and is defined by

$$\mathrm{EN}(S) = \ell - \ell_{\mathrm{c}}. \tag{3.5}$$

Rearranging expression (3.5) provides a decomposition of the log-likelihood ℓ in a complete log-likelihood term ℓ_{c} and the entropy $\mathrm{EN}(S)$. Moreover, the entropy $\mathrm{EN}(S)$ measures the difference between the maximum likelihood approach (ML) of the mixture model and the classification maximum approach (CML) when a model-based clustering is under consideration (see details in Celeux and Soromenho (1996, Section 3.1)). If the entropy $\mathrm{EN}(S)$ is close to zero, both approaches can be thought as equivalent and this occurs when the clusters are well separated. The entropy takes a large value if the mixture components are poorly separated. Thus, the use of the entropy function in the definition of ICL-BIC allows for comparing the ML and CML approaches and assessing mixture models to provide a model-based clustering.

Classification likelihood criterion (CLC)

The classification likelihood criterion (CLC) was introduced by Biernacki and Govaert (1997) and makes use of the complete and incomplete log-likelihood function association defined by equation (3.5). Its origin derives from the fitting of normal mixture models. The CLC is formulated as equation (3.3) with a penalty term $P = 2\mathrm{EN}(S)$.

This criterion does not include a complexity term due to the number of parameters. Therefore, it should allow less parsimonious models compared to other measures that tends to penalize against large number of parameters (i.e. complex models) such as AIC, BIC, and ICL-BIC. In addition, the CLC tends to penalize poorly separated clusters because it uses the entropy $\mathrm{EN}(S)$ of the fuzzy classification (as the ICL-BIC does).

Biernacki et al. (1999) stated that the CLC tends to overestimate the number of groups in the data compared to ICL-BIC and it only works well when the mixing proportion values are restricted to be equal.

Consistent Aikake Information Criterion (CAIC)

This criterion was proposed by Bozdogan (1987) as a variant of the AIC which makes the AIC asymptotically consistent. Its formulation from equation (3.3) is defining the penalty term as $P = K(1 + \log(n))$, i.e. larger penalty term than BIC and therefore even higher penalty against complex models.

One of the main criticisms of the CAIC is that it tends to select simpler models than AIC does. For example, Anderson et al. (1998) compared the performance of the AIC and CAIC in capture-recapture datasets concluding that the models selected by the CAIC are overly parsimonious and with poor structure and, therefore, they recommend the use of the AIC instead of the CAIC.

Normalized Entropy Criterion (NEC)

The normalized entropy criterion (NEC) was originally introduced by Celeux and Soromenho (1996). In a different approach, the NEC arises from the idea of using the entropy function alone as a criterion for choosing the number of clusters. For S clusters, this criterion is defined as follows:

$$\text{NEC}(S) = \frac{\text{EN}(S)}{\ell - \ell^{(1)}} = \frac{\ell - \ell_{\text{c}}}{\ell - \ell^{(1)}}, \tag{3.6}$$

where $\ell^{(1)}$ is the value of the maximised incomplete-data log-likelihood for a single ($S = 1$) component. As formulated in (3.6), NEC(1) is not defined and therefore this criterion suffers of limitation that it cannot choose between one and more than one cluster. Celeux and Soromenho (1996) proposed a rule of thumb for this case, but their procedure was restricted to normal mixtures.

Biernacki et al. (1999) found that the original NEC as defined in (3.6) performs in an unsatisfactory way and proposed a modification in its use. They referred to this new procedure as improved NEC criterion. Effectively, they defined $\text{NEC}(1) = 1$, i.e. define NEC to be one for $S = 1$ (see a brief review of the justification for this in McLachlan and Peel (2000, Section 6.10.2)). The improved NEC criterion simply then consists of choosing the number of groups S to minimize NEC(S). If all the S groups make $\text{NEC}(S) > 1$, then there is no cluster structure in the data and we choose $S = 1$. Otherwise, we choose the number of clusters S with minimum NEC value. Biernacki et al. (1999) showed through numerical examples that this improved procedure of the NEC criterion corrects for

the tendency of the original NEC to prefer $S > 1$ clusters when the true number is $S = 1$.

Approximate weight of evidence (AWE)

The approximate weight of evidence (AWE) was proposed by Banfield and Raftery (1993) as a Bayesian solution to the choice of the number of clusters using the complete data log-likelihood function. The AWE criterion is formulated as:

$$\text{AWE} = -2\ell_{\text{c}} + 2K\left(\frac{3}{2} + \log(n)\right) = \text{BIC} + 3K.$$

This criterion selects more parsimonious models than BIC because it penalizes more complex models (additional term $3K$). The drawback of this criterion is that parameter estimation is biased when the clusters are not well separated (see McLachlan and Peel (2000, Section 2.21) for details).

$\mathcal{L}$ criterion

$\mathcal{L}$ criterion was introduced by Figueredo and Jain (2002) as a result of a novel technique which is an alternative to the EM algorithm and might be applied to any type of parametric mixture model for which it is possible to write the EM algorithm. The $\mathcal{L}$ criterion has the form:

$$\mathcal{L} = -\ell - \frac{K}{2}\sum_{s=1}^{S}\log\left(\frac{n\pi_s}{12}\right) - \frac{S}{2\log\left(\frac{n}{12}\right)} - \frac{S(K+1)}{2}$$

where $\{\pi_s\}$ are the cluster membership probabilities.

A summary table with the definitions of all these information criteria measures are given in Table 3.1.

3.2.3 Information Criteria. Differences and Comparisons

A distinction among the information criteria presented in the previous section is regarding the aim for which the criterion was proposed. We can differentiate two groups here: a first group where the criteria were developed for regression models (AIC, AIC_{c}, AIC_{u}, AIC3, CAIC and BIC) and second group where the criteria were proposed for clustering (CLC, ICL-BIC, NEC, AWE and $\mathcal{L}$).

Table 3.1: Information criteria summary table for one-dimensional clustering case.

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

Notes: nm is the total sample size which is the number of elements in the response matrix Y. K is the number of parameters, S the number of clusters, π_S the mixing cluster proportion, ℓ the the maximised incomplete data log-likelihood, ℓ_c is the maximised complete data log-likelihood (see eq. (2.11) for row clustering and eq. (2.13) for column clustering). $\text{EN}(\cdot)$ is the entropy function defined by $\text{EN}(S) = \ell - \ell_\text{c}$.

Another difference is with respect to which parameters define each criterion apart from the maximised complete-data and incomplete-data log-likelihood. AIC and AIC3 depend on the number of free parameters K. AIC_c, AIC_u, CAIC and BIC are determined by both the sample size n and the number of free parameters K. NEC and CLC are defined by the entropy function. ICL-BIC and AWE depend on the sample size n, the number of free parameters K and the entropy function. Finally, the $\mathcal{L}$ criterion is defined by the sample size n, the number of free parameters K and the cluster membership probabilities $\{\pi_S\}$.

A final distinction is that BIC, AWE and ICL-BIC are criteria motivated from a

Bayesian perspective for choosing the number of components in a mixture model. There might be a Bayesian interpretation for the other information criteria, but it is questionable in many situations (Steele and Raftery, 2010).

There have been several simulation studies dealing with the empirical comparison of the information criteria performances for mixture models with continuous, categorical and mixed (continuous and categorical) distributions. For example, McLachlan and Ng (2000) brought into comparison AIC, BIC, CLC, ICL-BIC among others for three simulated data set of normal mixture components. The main conclusions are that ICL-BIC has good overall performance and the tendency of AIC to select too many normal components. Bezdek et al. (1997) showed experiments from bivariate normal distribution as mixture components. AIC3, AIC, BIC, AWE and NEC have the best results (in that order). Bozdogan (1994) simulated 3-component mixture models with three-dimensional multivariate normal distributions. AIC3 and CAIC have the best performance.

Fonseca and Cardoso (2007) and Fonseca (2008) reported experiment results for finite mixture with continuous, categorical and mixed components. The results are that BIC has the best performance for normal multivariate distributions, AIC3 for multinomial distributions and ICL-BIC for mixed distributions. They conclude that AIC, AIC3, AIC_c and AIC_u are sensitive to the type of outcome variable.

From a Bayesian perspective, Steele and Raftery (2010) showed a comprehensive simulation study for Gaussian mixture models based on reported parameter values in the literature from 43 papers. BIC has the best performance.

All these simulation experiments suggest that the type of outcome variable may affect the results of the information criterion performance. For that reason, in order to determine the best information criterion for our proposed likelihood-based methodology for ordinal data, we carry out a simulation study specific to our clustering of ordinal data.

3.3 Simulation Study

3.3.1 Methodology

In this section, we evaluate the performances of eleven of the most common information criterion measures with ordinal data and finite mixtures: AIC, AIC_c,

BIC, ICL-BIC, AIC_u, AIC3, CLC, CAIC, NEC, AWE and the $\mathcal{L}$ criterion. Their definitions are given in Table 3.1.

The goal of the experiments is to assess the performance of these information criteria in determining the true number of clusters. In particular, the results we are interested in are the percentage of total simulated experiments where each information criterion correctly determines the correct number of row/column clusters in a diverse set of scenarios. The scenarios are determined by varying the sample size/subjects ($n = 50, 100, 500$) and number of measures/questions ($m = 5, 10$). In addition, we made variations in the number of row clusters ($R = 2, 3, 4$), column clusters ($C = 2, 3, 4$) and the space between the $q = 4$ score parameters $\{\phi_k\}$. The experimental design for the row clustering and the biclustering cases is given in Tables 3.2 and 3.3, respectively.

Table 3.2: Factorial design for the simulation study. Row clustering.

	Number of Clusters $\boldsymbol{R}$	**Sample Sizes** $\boldsymbol{n}$	**Scenarios**	**Total**	**Replicates**
	2	50	Varying $m, \{\phi_k\}, \{\pi_r\}$	$\mathbf{3^2 \times 5}$	**100**
	3	100			
	4	500			
Levels	**3**	**3**	**5**	**45**	**4500 datasets**

Table 3.3: Factorial design for the simulation study. Biclustering.

	Number of Clusters $\boldsymbol{R}$	**Number of Clusters** $\boldsymbol{C}$	**Sample Sizes** $\boldsymbol{n}$	**Scenarios**	**Total**	**Replicates**
	2	2	50	Varying $m, \{\phi_k\}, \{\pi_r\}$	$\mathbf{2^2 \times 3 \times 5}$	**100**
	3	3	100			
			500			
Levels	**2**	**2**	**3**	**5**	**60**	**6000 datasets**

Through variations in the number of clusters, the number of measures/questions and the spacing between one pair of adjacent score parameters we have constructed five different scenarios. The scenarios tested illustrate a comprehensive set of situations representing from a simple to a more challenging context. The latter situation allows us to test information criteria in demanding situations where

the estimation of the true number of clusters might be expected to be more difficult. The five scenarios for $\{\phi_k\}$ may be described by: equal spacing between any pair of adjacent score parameters (Scenario 1), one pair of adjacent score parameters are very close in value (Scenario 2), one of the mixing cluster proportions is close to zero (Scenario 3), one pair of adjacent score parameters have the same value (Scenario 4), and the same as the first scenario but increasing the number of measures to $m = 10$ (Scenario 5). Table 3.4 shows the parameter configuration regarding the mixing proportions and the score parameters in the row clustering case. All the parameters for each scenario in the row clustering and biclustering cases are given in Tables B.1 and B.2 in Appendix B.1.

Table 3.4: Row clustering. Row membership probabilities $\{\pi_r\}$ and score parameters $\{\phi_k\}$ for 5 tested scenarios. The value of the other parameters are shown in Table B.1 in Appendix B.1. The aspects of each scenario which we expect to be challenging are coloured blue.

	Scenario 1 $m = 5$	**Scenario 2** $m = 5$	**Scenario 3** $m = 5$	**Scenario 4** $m = 5$	**Scenario 5** $m = 10$
$R = 2$	$\pi_1 = 0.45$	$\pi_1 = 0.45$	$\pi_1 = 0.95$	$\pi_1 = 0.45$	$\pi_1 = 0.45$
	$\pi_2 = 0.55$	$\pi_2 = 0.55$	$\boldsymbol{\pi_2 = 0.05}$	$\pi_2 = 0.55$	$\pi_2 = 0.55$
	$\phi_2 = 0.34$	$\phi_2 = 0.34$	$\phi_2 = 0.34$	$\boldsymbol{\phi_2 = 0.5}$	$\phi_2 = 0.34$
	$\phi_3 = 0.66$	$\boldsymbol{\phi_3 = 0.97}$	$\phi_3 = 0.66$	$\boldsymbol{\phi_3 = 0.5}$	$\phi_3 = 0.66$
$R = 3$	$\pi_1 = 0.20$	$\pi_1 = 0.20$	$\pi_1 = 0.47$	$\pi_1 = 0.20$	$\pi_1 = 0.20$
	$\pi_2 = 0.50$	$\pi_2 = 0.50$	$\boldsymbol{\pi_2 = 0.05}$	$\pi_2 = 0.50$	$\pi_2 = 0.50$
	$\pi_3 = 0.30$	$\pi_3 = 0.30$	$\pi_3 = 0.48$	$\pi_3 = 0.30$	$\pi_3 = 0.30$
	$\phi_2 = 0.34$	$\phi_2 = 0.34$	$\phi_2 = 0.34$	$\boldsymbol{\phi_2 = 0.5}$	$\phi_2 = 0.34$
	$\phi_3 = 0.66$	$\boldsymbol{\phi_3 = 0.97}$	$\phi_3 = 0.66$	$\boldsymbol{\phi_3 = 0.5}$	$\phi_3 = 0.66$
$R = 4$	$\pi_1 = 0.15$	$\pi_1 = 0.15$	$\pi_1 = 0.31$	$\pi_1 = 0.15$	$\pi_1 = 0.15$
	$\pi_2 = 0.30$	$\pi_2 = 0.30$	$\boldsymbol{\pi_2 = 0.05}$	$\pi_2 = 0.30$	$\pi_2 = 0.30$
	$\pi_3 = 0.25$	$\pi_3 = 0.25$	$\pi_3 = 0.32$	$\pi_3 = 0.25$	$\pi_3 = 0.25$
	$\pi_4 = 0.30$	$\pi_4 = 0.30$	$\pi_4 = 0.32$	$\pi_4 = 0.30$	$\pi_4 = 0.30$
	$\phi_2 = 0.34$	$\phi_2 = 0.34$	$\phi_2 = 0.34$	$\boldsymbol{\phi_2 = 0.5}$	$\phi_2 = 0.34$
	$\phi_3 = 0.66$	$\boldsymbol{\phi_3 = 0.97}$	$\phi_3 = 0.66$	$\boldsymbol{\phi_3 = 0.5}$	$\phi_3 = 0.66$

Notes: $\phi_1 = 0$ and $\phi_4 = 1$ for all the scenarios.
Scenarios 1-4 have 11 free parameters and scenario 5 has 16.

In the case of only clustering the rows, the simulation study is tested in 45 different combinations (see Table 3.2) over 5 scenarios. For each scenario, we drew $h = 100$ datasets, and select the best model for each dataset using each information criterion. Therefore, we worked with 4500 $(9 \times 5 \times 100)$ samples. In

the same manner, the number of samples generated in the case of biclustering were 6000 $(12 \times 5 \times 100)$.

The EM algorithm to obtain the estimates is repeated 10 times with random starting points and the estimates with the highest likelihood are kept. This helps to avoid local optima in the estimation process because the log-likelihood function has a multimodal surface.

A sketch of the simulation study procedure is given in the following section.

3.3.2 Simulation Study Outline

The simulation study procedure for the row clustering case is outlined in the following steps:

Step 1. Model specification

Select the model, w, from a set of models $w = 1, \ldots, W$.
There are $W = 3 \times 1 \times 5$ possible models:

- Select $R \in \{2, 3, 4\}$ (3 options).
  This fixes $\{\alpha_1, \ldots, \alpha_R\}$ (with $\sum_{r=1}^{R} \alpha_r = 0$).
- Set the number of response categories: $q = 4$ in all cases (1 option).
  This fixes $\{\mu_1, \ldots, \mu_q\}$ (with $\mu_1 = 0$).
- Select a scenario from one of the five in a set of predefined scenarios (5 options).
  Each scenario fixes:
  - Prior mixing probabilities $\pi_1, \ldots, \pi_R$ (with $\sum_{r=1}^{R} \pi_r = 1$).
  - The number of columns $m \in \{5, 10\}$.
    This fixes $\{\beta_1, \ldots, \beta_m\}$ (with $\sum_{j=1}^{m} \beta_j = 0$).
  - The ordinal response cut levels $\phi_1 \leq \phi_2 \leq \ldots \leq \phi_q$ (with $\phi_1 = 0$ and $\phi_q = 1$).

At the end of this step we know, for the chosen model w:

- The number of row groups R^w.
- The number of response categories q^w.

- The number of columns m^w.
- The total number of free parameters K^w.
- The parameter values:

$$\{\alpha_1^w, \ldots, \alpha_R^w\}, \{\beta_1^w, \ldots, \beta_m^w\}, \{\pi_1^w, \ldots, \pi_R^w\}, \{\mu_1^w, \ldots, \mu_q^w\}, \{\phi_1^w, \ldots, \phi_q^w\}$$

  and as a consequence we can calculate the values of the linear predictors

$$\eta_{krj}^w = \mu_k^w + \phi_k^w \left(\alpha_r^w + \beta_j^w\right)$$

  for $k \in \{1, \ldots, q^w\}$, $r \in \{1, \ldots, R^w\}$ and $j \in \{1, \ldots, m^w\}$.

Step 2. Sample size specification

Select the sample size label, s, from a set of possible labels $s = 1, \ldots, S$.
There are $S = 3$ possible sample sizes, $s \in \{1, 2, 3\}$:

- Select $n_s \in \{50, 200, 500\}$ (3 options).

There are $WS = 15 \times 3 = 45$ possible combinations of model and sample size: (ws).

Step 3. Generate replicate datasets

There are $H = 100$ replicates.
For each model w and sample size s and each replicate $h \in \{1, \ldots, H\}$:

- For each row $i = 1, \ldots, n_s$, generate row membership

$$\mathbf{z}_i^{wsh} = \left(Z_{i1}^{wsh}, ..., Z_{iR}^{wsh}\right) \sim \text{Multinomial}\left(1; \{\pi_r^w\}\right).$$

- For each column $j = 1, \ldots, m^w$ within each row $i = 1, \ldots, n_s$, generate the response ordinal variable

$$y_{ij}^{wsh} | \mathbf{z}_i^{wsh} = \boldsymbol{\delta}_r \sim \text{Stereotype}\left(\left\{\eta_{krj}^w\right\}_{k=1}^q\right).$$

Here $\boldsymbol{\delta}_r$ is an indicator vector of length R^w, with 1 at location r and zero elsewhere. This implies that

$$\log\left(\frac{\mathrm{P}\left[y_{ij}^{wsh}=k \mid \mathbf{z}_i^{wsh}=\boldsymbol{\delta}_r\right]}{\mathrm{P}\left[y_{ij}^{wsh}=1 \mid \mathbf{z}_i^{wsh}=\boldsymbol{\delta}_r\right]}\right)=\eta_{krj}^{w}.$$

There are $WSH = 15 \times 3 \times 100 = 4500$ possible combinations of model, sample size, and replicate: (wsh).

Step 4. Fit models

We fit models with $r = 1, \ldots, R_{\max}$ row groups to each dataset with $R_{\max} = 8$.

- For each r we run the EM algorithm $F = 10$ times. On run f we randomly generate a starting point $\Omega_{rf}^{(0)}$ by drawing the parameter values independently from the following distributions,

$$\begin{aligned}
\mu_k^{(0)} &\sim \text{Uniform}(-5,5) && k=2,\ldots,q,\\
\alpha_\ell^{(0)} &\sim \text{Uniform}(-5,5) && \ell=1,\ldots,r-1,\\
\beta_j^{(0)} &\sim \text{Uniform}(-5,5) && j=1,\ldots,m-1,\\
\phi_k^{(0)} &\sim \text{Uniform}(\phi_{k-1}^{(0)},1) && k=2,\ldots,q-1,\\
\left(\pi_1^{(0)},\ldots,\pi_r^{(0)}\right) &\sim \text{Dirichlet}\left(1;\lambda_1=1,\ldots,\lambda_r=1\right).
\end{aligned}$$

  This means running the EM algorithm $R_{\max} \times F = 8 \times 10 = 80$ times on each of the 4500 datasets, a total of 36000 runs in all.

  On run f for row group number r fitted to replicate dataset wsh we obtain parameter estimate $\widehat{\Omega}_{rf}^{wsh}$, each with its associated complete log-likelihood value $\ell_c^w(\widehat{\Omega}_{rf}^{wsh})$.

- For each row group number r fitted to the dataset wsh we select the parameter estimate $\widehat{\Omega}_{r*}^{wsh}$ as the one which has the largest associated complete data log-likelihood $\{\ell_c^w(\widehat{\Omega}_{rf}^{wsh})\}_{f=1}^{F}$. This is the best fit to the dataset wsh with r row groups.

- Next we calculate a value for each of the L information criteria at each of

these best fit values:

$$C_{r\ell}^{wsh} = \text{CRI}_\ell(\ell_c^w(\widehat{\Omega}_{r*}^{wsh}), \ell^w(\widehat{\Omega}_{r*}^{wsh}), n_s, m^w, K^w, r, \{\pi_r\}) \qquad \text{for } \ell = 1, \ldots, L$$

Note that the information criteria can depend on the complete and incomplete data log-likelihood value, the sample size n_s, the number of columns m^w, the number of fitted parameters K^w, the number of fitted row groups r, and the mixing proportions π_r.

- For each criterion we identify over the values of r the minimum value of $C_{r\ell}^{wsh}$, and the corresponding number of row groups r for which that occurs:

$$r_{\ell*}^{wsh} = \underset{r}{\text{argmin}} \; C_{r\ell}^{wsh}$$

This is the number of row groups selected by criterion CRI_ℓ for the dataset wsh.

- The proportion of times across the H replicates generated by the same model w with the same sample size n_s where this selected number of row groups agrees with the true value R^w is of primary interest

$$P_\ell^{ws} = \frac{1}{H} \sum_{h=1}^{H} I(r_{\ell*}^{wsh} = R^w)$$

The best performing criterion ℓ is the one where P_ℓ^{ws} is consistently large, over a wide range of scenarios w and sample sizes n_s.

We are also interested in the mean, median and interquartile range of the values $\{r_{\ell*}^{wsh}\}_{h=1}^{H}$, to see which of these may consistently over or underestimate the number of row groups R.

This simulation study procedure refers to the one-dimensional clustering case and has been illustrated with the row clustering version. The simulation study outline for the column clustering version is basically the same just replacing parameters related to rows with the equivalent column parameters (e.g. the column mixing probabilities $\{\kappa_c\}$ instead of the row ones $\{\pi_r\}$). For the case of biclustering, we have evaluated the performance of the same information criterion measures as in the one-dimensional case (see Table 3.1). However, the asymptotic

properties in the information criterion measures used in this chapter might not apply in the case of biclustering. These properties apply to one-dimensional (e.g. number of rows n) assuming that the other dimension (i.e. number of columns, m) is fixed and therefore there are not asymptotic properties in m. Thus, the information criteria affects the two clusterings differently and it would be a future research direction to explore.

3.3.3 Results

Figure 3.1 is a histogram displaying the percentage of cases in which each information criterion determines the true number of row clusters pooling all results across the five scenarios and the factors used in the experimental control. Its equivalent histogram for biclustering is given in Figure 3.2. For row clustering, the overall best performance was AIC (correctly selecting the number of row clusters in 93.8% of cases), followed by AIC_c (89.8%) and AIC_u (82.4%). In the case of biclustering, the results are very similar as AIC also performs the best, although with a lower percentage of correctly selecting the number of row and column clusters than the row clustering case (86.1%). AIC_c and AIC_u also perform very well with percentages close to AIC: 85.6% and 84.2% respectively.

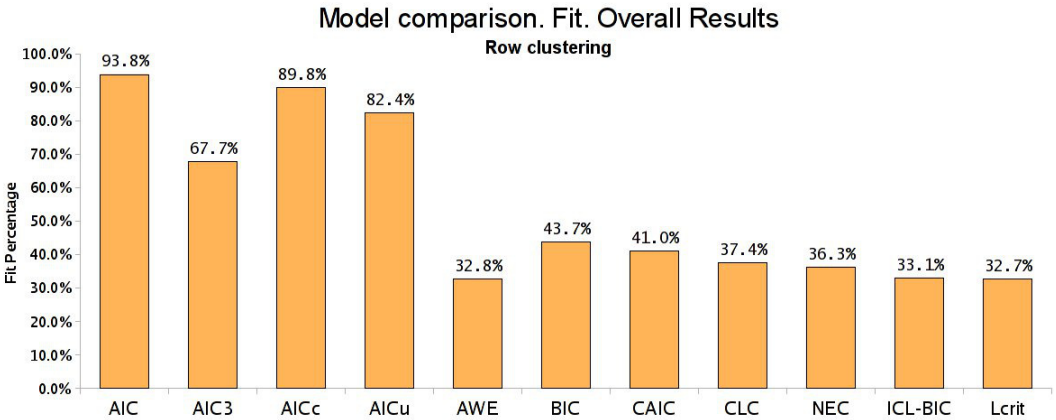

Figure 3.1: ***Simulation study results for row clustering:*** *Bars depict the percentage of cases for each information criterion correctly fits the true number of row clusters.*

Tables 3.5 and 3.6 show the best 5 information criterion performances over the 5 scenarios in the case of row clustering and biclustering respectively. In both cases, we can observe that the ranking is exactly the same over the 5 scenarios: AIC, AIC_c, AIC_u, AIC3 and BIC. The best performance is the scenario 5 which

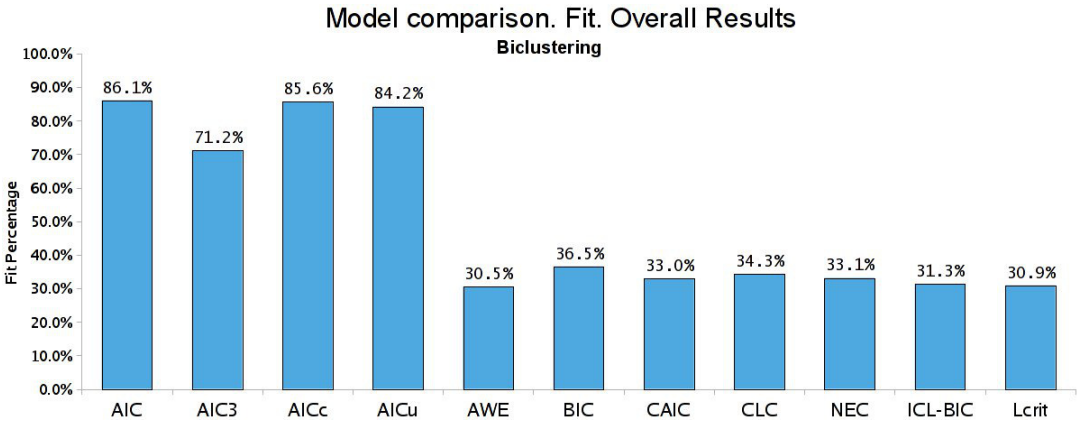

Figure 3.2: ***Simulation study results for biclustering:*** *Bars depict the percentage of cases for each information criterion correctly fits the true number of clusters.*

Table 3.5: Model comparison simulation study results. Row clustering. Ranking of the best 5 information criterion measures.

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

Table 3.6: Model comparison simulation study results. Biclustering. Ranking of the best 5 information criterion measures.

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

has the largest number of measures/questions m. On the other hand, the worst achievement is in the challenging scenario 3 (one of the mixing cluster proportions is close to zero). Regardless of the difficulty of this scenario, AIC and AIC_c performances are still quite satisfactory (above 75% of accurately choosing the correct number of clusters in row clustering and biclustering).

The full results for all the scenarios broken down by number of row/column clusters and the sample size is given from Table B.3 to Table B.12 in Appendix

B.2. A summary of those results broken down by scenario in the case of row clustering and biclustering respectively are given in Table 3.7 and Table 3.8. BIC is underestimating the number of clusters (incorrectly selecting a smaller number of clusters in 56% and 63.2% of cases in row clustering and biclustering respectively). The formulation of the penalty term in BIC allows us to penalize more complex models than AIC does. The results for CAIC are very similar (underestimating 58.7% and 66.9% of the cases) as the penalty term is similar to BIC and, therefore, the penalization is decreasing its performance. AWE obtains even worse results than BIC and CAIC (underestimating 64.2% and 63.1% of cases). This result is expected because the penalty term for AWE penalizes even more complex models than BIC ($P = \frac{3}{2}\log(n) > \log(n)$, as $n \geq 1$).

Very poor performance is obtained with ICL-BIC (correctly selecting the number of clusters in only 33.1% and 31.3% of cases in row clustering and biclustering respectively). Our results are in accordance with Fonseca and Cardoso (2007) for the categorical case (their results were correctly selecting the number of row clusters in only 16% of cases). ICL-BIC is only working correctly when the true number of row clusters is $R = 2$ (see the corresponding results in Tables B.3-B.12), so that ICL-BIC only identifies that there is cluster structure in the data. However, ICL-BIC generally underestimates the number of clusters (65.9% vs. 57.6%) probably due to its link with BIC (ICL-BIC is defined as BIC plus the entropy function).

CLC only performs well when the unknown mixing proportions are restricted to be very similar (Biernacki et al., 1999). For example, scenarios 1, 2, 4 and 5 have $R = 2$ with $\pi_1 = 0.45$ and $\pi_2 = 0.55$ (Tables B.3-B.4 and B.6-B.7 in the Appendix B.2). In those cases, there CLC performs very well (above 96.3%) in comparison with $\pi_1 = 0.95$ and $\pi_2 = 0.05$ (Scenario 3, Table B.5 in the Appendix B.2) where the percentage of correctly selecting the number of clusters drops to 85.3%. However, the overall CLC performance only correctly selects the number of clusters in 37.4% (row clustering) and 34.3% (biclustering) of cases. Finally, NEC tends to overestimate the number of clusters more than the other criteria (overestimating in 15.8% and 54.9% of the cases in row clustering and biclustering respectively) and the $\mathcal{L}$ criterion obtains similar results to ICL-BIC.

Based on our simulation study we therefore conclude that AIC is the best information criterion when dealing with ordinal data and we fit likelihood-based finite mixture models with the ordinal stereotype model as the components in the

mixture.

Table 3.7: Model comparison simulation study. Overall results for 11 information criteria over 5 scenarios. Row Clustering ($\mu_k + \phi_k(\alpha_r + \beta_j)$).

Results		AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall	Underfit	5.1	31.6	9.0	16.4	64.2	56.0	58.7	61.2	47.9	65.9	65.9
	Fit	93.8	67.7	89.8	82.4	32.8	43.7	41.0	37.4	36.3	33.1	32.7
	Overfit	1.1	0.7	1.2	1.2	0.8	0.3	0.3	1.3	15.8	1.1	1.4
Scenario 1	Underfit	7.9	37.7	9.1	20.3	66.4	57.9	62.8	62.7	44.3	66.0	66.1
	Fit	91.4	61.7	90.2	79.0	32.4	41.2	36.1	36.1	36.4	33.1	32.9
	Overfit	0.7	0.7	0.7	0.7	1.1	0.9	1.1	1.2	19.2	0.9	1.0
Scenario 2	Underfit	2.0	34.2	4.9	19.8	66.4	60.9	63.3	62.2	51.8	66.0	66.4
	Fit	97.6	65.6	94.8	80.0	33.6	39.1	36.7	37.8	36.0	34.0	32.9
	Overfit	0.4	0.2	0.3	0.2	0.0	0.0	0.0	0.0	12.2	0.0	0.7
Scenario 3	Underfit	10.0	42.4	23.3	31.3	66.0	60.0	62.7	60.2	42.0	65.3	65.1
	Fit	88.0	56.7	74.7	66.7	31.1	40.0	37.3	34.9	38.7	30.2	31.1
	Overfit	2.0	0.9	2.0	2.0	2.9	0.0	0.0	4.9	19.3	4.4	3.8
Scenario 4	Underfit	5.8	43.1	7.6	10.7	66.7	60.2	62.9	61.3	45.3	66.2	65.6
	Fit	92.9	56.4	91.1	88.0	33.3	39.6	37.1	38.4	36.7	33.8	33.6
	Overfit	1.3	0.4	1.3	1.3	0.0	0.2	0.0	0.2	18.0	0.0	0.9
Scenario 5	Underfit	0.0	0.4	0.0	0.0	55.6	40.9	42.0	59.8	56.2	65.8	66.2
	Fit	99.1	98.2	98.2	98.2	33.3	58.7	57.6	40.0	33.8	34.2	32.9
	Overfit	0.9	1.3	1.8	1.8	0.0	0.4	0.4	0.2	10.0	0.0	0.9

Notes: All the data is shown in percentage form(%).

Table 3.8: Model comparison simulation study. Overall results for 11 information criteria over 5 scenarios. Biclustering ($\mu_k + \phi_k(\alpha_r + \beta_c)$).

Results		AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall	Underfit	11.7	27.4	12.3	13.9	63.1	63.2	66.9	48.5	12.0	57.6	61.7
	Fit	86.1	71.2	85.6	84.2	30.5	36.5	33.0	34.3	33.1	31.3	30.9
	Overfit	2.2	1.4	2.1	1.9	6.4	0.2	0.2	17.2	54.9	11.1	7.5
Scenario 1	Underfit	9.5	23.5	9.7	14.0	64.2	65.3	68.8	50.0	15.2	58.2	61.7
	Fit	89.2	75.8	89.2	84.8	28.5	34.5	31.0	29.5	42.5	29.2	31.3
	Overfit	1.3	0.7	1.2	1.2	7.3	0.2	0.2	20.5	42.3	12.7	7.0
Scenario 2	Underfit	16.5	33.7	17.3	18.2	61.8	64.8	68.2	52.3	15.2	57.7	58.8
	Fit	82.3	65.5	81.5	80.7	31.8	35.2	31.8	31.2	35.7	32.0	32.5
	Overfit	1.2	0.8	1.2	1.2	6.3	0.0	0.0	16.5	49.2	10.3	8.7
Scenario 3	Underfit	16.7	33.5	17.3	18.5	63.2	66.3	69.2	49.8	15.0	57.0	60.5
	Fit	80.5	64.7	80.0	79.3	30.3	33.5	30.7	33.8	40.8	31.2	30.5
	Overfit	2.8	1.8	2.7	2.2	6.5	0.2	0.2	16.3	44.2	11.8	9.0
Scenario 4	Underfit	13.3	28.3	14.5	15.8	64.3	67.5	70.2	50.2	13.2	59.3	64.2
	Fit	85.5	70.8	84.5	83.3	28.2	32.3	29.7	30.5	35.0	27.3	29.2
	Overfit	1.2	0.8	1.0	0.8	7.5	0.2	0.2	19.3	51.8	13.3	6.7
Scenario 5	Underfit	2.7	13.7	2.8	2.8	62.1	52.2	57.9	40.1	1.5	55.8	63.2
	Fit	92.8	83.3	92.8	92.8	33.4	47.2	41.8	46.6	11.3	36.9	30.9
	Overfit	4.5	3.0	4.3	4.3	4.5	0.7	0.3	13.3	87.2	7.3	5.9

Notes: All the data is shown in percentage form(%).

3.4 Application to Real Data with Known R

In this section, we use the best 5 information criteria according to the results of the simulation study (AIC, AIC_c, AIC_u, AIC3 and BIC) to select the fitted finite mixture model which best represents a real-life ordinal dataset. The set of candidate models is estimated with our likelihood-based clustering approach.

The real-life example we analyse is an ordinal dataset from community psychology collected by Anders and Batchelder (2013) where they artificially created three different response profiles among the participants. The data are the responses of 83 respondents to 20 questions about a particular city. In order to create the three distinct cultures, the respondents all received the same city-knowledge questionnaire but were randomly assigned to answer the questions regarding one of the following three cities: Irvine, California; New York, New York; or Miami, Florida. Thus, 30 respondents answered for Irvine, 29 for Miami, and 24 for New York. All the questions are related to magnitudes (e.g. amount of rain, snow, or level of humidity in the city) and are categorised on an ordinal 7-point scale, from low to high magnitude. Tables B.13 and B.14 in Appendix B.3 shows the full set of questions and the whole data set respectively. The choice of those three particular cities for the study was because they are well distinguished in terms of the expected responses to the questionnaire. Therefore, the aim of the study was to deliberately design an experiment that would encourage three distinct cultural profiles, one for each city.

Our main goal is to apply our likelihood-based clustering approach to this dataset in order to identify that the model which best represents the data includes $R = 3$ respondent (row) clusters based on their responses over the 20 questions. The cluster structure should identify the three distinct response profiles. We have fitted a suite of row clustering models based on our methodology, from row clustering without column effects model to row clustering with and without an interaction. For each model, the 5 best-performing information criteria for the row clustering case according to the simulation study were computed and the results are given in Table 3.9.

The five information criteria indicate that the best model is the stereotype model including row clustering with $R = 3$ row (respondent) groups and with interaction factors ($\mu_k+\phi_k(\alpha_r+\beta_j+\gamma_{rj})$) with AIC=5110.62, AIC_c=5117.24, AIC_u=5149.69, AIC3=5182.62 and BIC=5500.47. It is also remarkable that $R = 3$ respondent clus-

Table 3.9: Suite of row clustering models fitted for Anders and Batchelder (2013) data set. The calculation of the 5 best-performance information criteria for our approach is given for each model. For each information criterion, the best model in each category (row clustering without column effects, row clustering with and without interactions) is shown in boldface and the overall best model in blue boldface.

Model		R	C	npar	AIC	AIC_c	AIC_c	AIC3	BIC
Row clustering	$\mu_k + \phi_k\alpha_r$ (no column effects)	2	1	13	6155.33	6155.55	6161.66	6168.33	6225.72
		3	1	15	6159.91	6160.20	6167.19	6174.91	6241.13
		4	1	17	6106.88	6107.26	6115.12	6123.88	**6198.93**
		5	**1**	**19**	**6104.71**	**6105.17**	**6113.91**	**6123.71**	6207.58
	$\mu_k + \phi_k(\alpha_r + \beta_j)$ (no interaction)	2	m	32	5730.54	5731.84	5746.32	5762.54	5903.81
		3	$\boldsymbol{m}$	**34**	**5700.53**	**5701.99**	**5717.36**	**5734.53**	**5884.62**
		4	m	36	5709.60	5711.24	5727.50	5745.60	5904.52
		5	m	38	5722.73	5724.56	5741.71	5760.73	5928.48
	$\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj})$ (interaction)	2	m	51	5362.58	5365.88	5388.84	5413.58	5638.73
		3	$\boldsymbol{m}$	**72**	**5110.62**	**5117.24**	**5149.69**	**5182.62**	**5500.47**
		4	m	93	5387.90	5399.06	5441.13	5480.90	5891.45
		5	m	114	5423.60	5440.57	5492.40	6040.86	6040.86

ters is also the best model in the case of row clustering model without interaction factors. However, the row clustering model without column effects is too simple, and does not identify the $R = 3$ cluster structure as the best model.

Figure 3.3 depicts a plot with the average of the fitted scores of respondent answers over the 20 city-knowledge questions where each respondent is allocated to the row group to which the respondent belongs with highest posterior probability. The computation of this average of the fitted scores is developed in the Section 4.2.1. The row cluster to which the respondent is allocated according to the row clustering model with $R = 3$ row groups and interaction factors are indicated by points of different shapes and colours. The figure shows three clearly distinguished clusters which predominantly correspond to the three intentionally created city groups. According to the results of Anders and Batchelder (2013), we might name the three clusters as: "New York" cluster (green triangles), 'Miami" cluster (red circles) and "Irvine" cluster (black squares). Likewise, Figure 3.4 shows the histogram of the number of respondents by their average fitted score interval. The colours for each bar are related to the colour groups shown in Figure 3.3. The histogram clearly shows the same three distinct respondent groups. There is a single bar per each group from the histogram. That is due to the fact that fitted scores have very small variability within each group. It might be observed in Figure 3.3 since the three clouds of points are quite well separated.

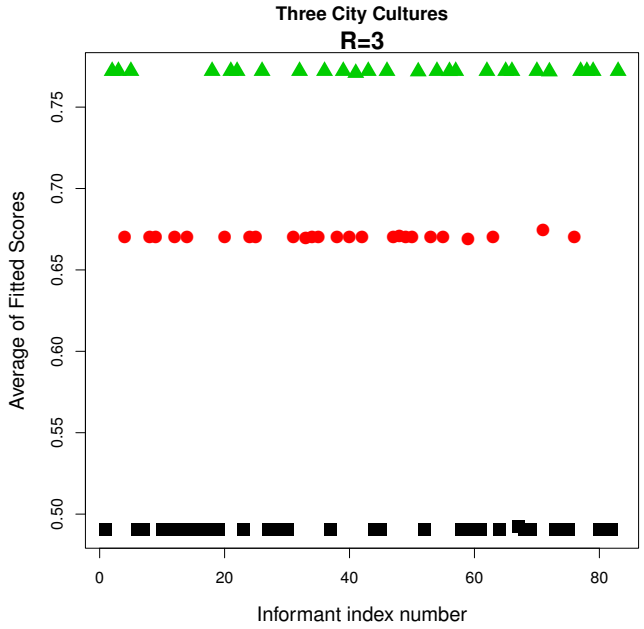

Figure 3.3: ***Anders and Batchelder (2013) data set:*** *The x-axis depicts the respondent index number (not informative). The y-axis depicts the average of the $R = 3$ fitted respondent clusters $\{\overline{\phi}_{(i.)}\}$ (see eq. (4.3)) from the row clustering version model with interaction factors.*

Finally, Table 3.10 shows the comparison between the original city membership allocation of the respondents and the clustering membership allocation from our methodology. In the latter, each respondent is assigned to the row cluster to which it is most likely to belong a posteriori. The fuzzy clustering obtains general 3 groups that mostly match the original survey they were given (the percentage of correctly clustering the respondent to the original allocation city is 92.7%). Note that 6 respondents are incorrectly allocated. However, these results are in accordance with Anders and Batchelder (2013) clustering results and, therefore, those respondents are appropriately clustered into a different city group. They may have not followed directions (e.g. they responded their home city as Irvine when given a Miami questionnaire), or some respondents may have unrealistic ideas of the city to which they were assigned.

The conclusion is that all the best 5 information criteria according to the results of the simulation study perform correctly when an ordinal dataset is analysed fitting our likelihood-based clustering approach.

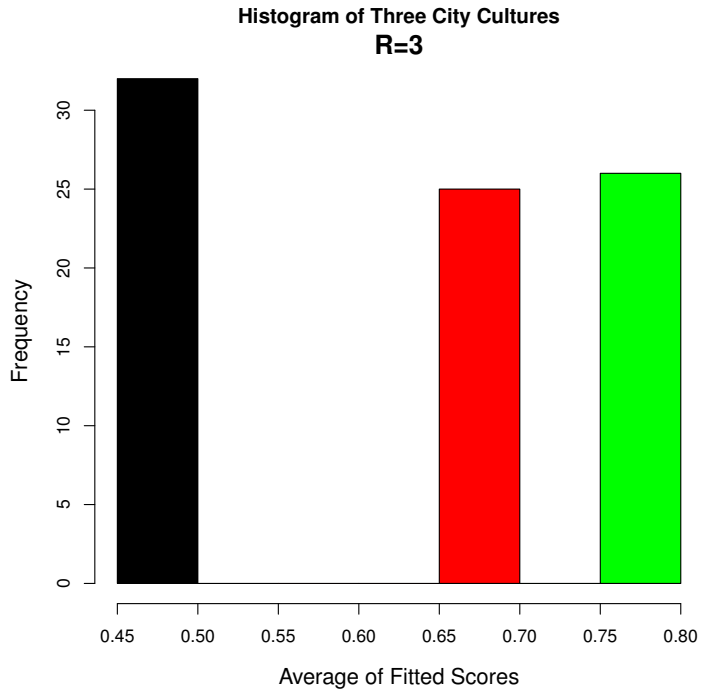

Figure 3.4: ***Anders and Batchelder (2013) data set:*** *Histogram of the $R = 3$ fitted respondent clusters $\{\overline{\phi}_{(i.)}\}$ (see eq. (4.3)) from the row clustering version model with interaction factors.*

3.5 Discussion

The results of our empirical study show that AIC, AIC_c and AIC_u are reliable information criteria to score fitted models based on our likelihood-based finite mixture model approach for ordinal datasets. In particular, AIC is the best information criterion for selecting the model with the correct number of clusters in a wide range of scenarios. It correctly selects the number of clusters in 93.8% (row clustering) and 86.1% (biclustering) of cases.

According to the results of our experiments, the best 5 information criterion performances are from AIC, AIC_c, AIC_u, AIC3 and BIC (see Tables 3.5 and 3.6). In order to test them in a real-life example, we used them to select the best fitted mixture model in a community psychology dataset. The analysed dataset was artificially generated with a known number of clusters and the aim was to confirm if those information criteria would correctly select the number of clusters and correctly classify each respondent. The results of fitting our approach to this dataset are very satisfactory because the selected estimated model identifies the

Table 3.10: Prediction of the clustering allocation ("Clustering") in comparison with the original allocation ("Original"). The true city is shown in row "True". "I" stands for Irvine, "N" for New York and "M" for Miami. The wrong allocations are shown in blue boldface.

Informant ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Original	I	N	N	M	N	I	I	M	M	I	I	M	I	M	I	I	**M**	N	I	M	N	N	I	M
Clustering	I	N	N	M	N	I	I	M	M	I	I	M	I	M	I	I	**I**	N	I	M	N	N	I	M

Informant ID	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Original	M	N	I	I	I	I	M	N	M	M	M	N	I	M	N	M	**M**	M	N	I	I	N	M	M
Clustering	M	N	I	I	I	I	M	N	M	M	M	N	I	M	N	M	**N**	M	N	I	I	N	M	M

Informant ID	**49**	50	**51**	**52**	53	54	55	56	57	58	59	60	61	62	63	64	65	66	**67**	68	69	70	71	72
Original	**I**	M	**M**	**M**	M	N	M	N	N	I	M	I	I	N	M	I	N	N	**M**	I	I	N	M	N
Clustering	**M**	M	**N**	**I**	M	N	M	N	N	I	M	I	I	N	M	I	N	N	**I**	I	I	N	M	N

Informant ID	73	74	75	76	77	78	79	80	81	82	83
Original	I	I	I	M	N	N	N	I	I	I	N
Clustering	I	I	I	M	N	N	N	I	I	I	N

correct (known) number of clusters. In addition, the clustering membership allocation resulting from our fuzzy clustering methodology concurs with the original membership allocation.

The simulation study is empirical and the conclusions are based on a set of information criteria in common use, none of which were developed for ordinal data. Furthermore, there has so far been minimal research on model selection for finite mixture models with categorical data. Because of this, development of a specific measure for model comparison with ordinal variables is required and should be achieved in future research. Additionally, as we indicated in Section 3.3.2, asymptotic properties in the information criterion measures for the two clusterings case may also be a future research direction to explore.

Chapter 4

Data Applications

In this chapter, the reliability of estimation of the stereotype model parameters is demonstrated in a simulation study (Section 4.1). In addition, we illustrate the stereotype model and our likelihood-based clustering method with three real-life examples (Section 4.2). In order to do model selection, the two best information criteria (AIC and AIC_c) according to the comparison study (see Chapter 3) are computed together with the most commonly used BIC and ICL-BIC. Thus, their performances can be compared.

4.1 Stereotype Models. Simulation Study

We set up a simulation study to test how reliably, in a diverse range of scenarios, we were able to estimate the parameters of stereotype models using the EM algorithm. We are not testing model selection here (that was tested in Chapter 3): instead we simulate datasets and then fit the correct model to those data.

The design of the simulation study includes an ordinal response variable with four categories and we varied the sample size ($n = 25, 50, 200, 500, 1000, 5000$), the number of columns ($m = 5, 10, 15$) and the number of row and/or column clusters ($R = 2, 3, 4, 5, 6$ and $C = 2, 3$). For each combination of sample size and number of row clusters, a single set of parameters values was chosen and 100 data sets (replicates) were generated. The MLEs and their standard errors were found for each replicate. The general results for the score parameters $\{\widehat{\phi}_k\}$ for row clustering are given in Table 4.1. Table 4.2 and Table 4.3 present the equivalent results for column clustering and biclustering respectively. The simulation

scenarios including the interaction factors for row clustering and biclustering version are showed in Tables C.1 and C.2 in Appendix C.1. In each case the tables show the mean of the MLEs and of their corresponding standard errors over the 100 replicates.

Table 4.1: Simulation study. Estimated score and row membership parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The number of categories is $q = 4$. MLEs and their standard errors from the score and row membership parameters ($\{\phi_k\}$,$\{\pi_r\}$) for different number of row clusters R and sample sizes n are shown.

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

For all models (row clustering, column clustering and biclustering) the estimates of the parameters $\{\phi_k\}$, $\{\pi_r\}$ and $\{\kappa_c\}$ are close to their true values and as expected the variability decreases with increasing the sample size n, and the number of columns m in the case of column clustering. Figure 4.1 shows the 100 separate estimates of $\widehat{\phi}_2$ and $\widehat{\phi}_3$ for the row clustering model with $R = 2$ row clusters plotted against each other for varying sample sizes. Note that all the es-

Table 4.2: Simulation study. Estimated score and column membership parameters for stereotype model including column clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_i} + \boldsymbol{\beta_c})$. The number of categories is $q = 4$. MLEs and their standard errors from the score and column membership parameters ($\{\phi_k\}$,$\{\kappa_c\}$) for different number of column clusters C, number of columns m and sample sizes n are shown.

C	Numpar	True param.	n=25					
			m=5		m=10		m=15	
			Mean	S.E.	Mean	S.E.	Mean	S.E.
2	31	$\phi_2 = 0.335$	0.291	0.261	0.314	0.143	0.329	0.100
		$\phi_3 = 0.672$	0.722	0.245	0.652	0.169	0.681	0.103
		$\kappa_1 = 0.600$	0.589	0.190	0.589	0.122	0.588	0.095
3	33	$\phi_2 = 0.335$	0.296	0.259	0.307	0.158	0.342	0.090
		$\phi_3 = 0.672$	0.790	0.283	0.712	0.177	0.682	0.110
		$\kappa_1 = 0.400$	0.371	0.204	0.376	0.124	0.390	0.087
		$\kappa_2 = 0.200$	0.179	0.196	0.195	0.112	0.195	0.086

C	Numpar	True param.	n=50					
			m=5		m=10		m=15	
			Mean	S.E.	Mean	S.E.	Mean	S.E.
2	56	$\phi_2 = 0.335$	0.397	0.215	0.348	0.119	0.335	0.081
		$\phi_3 = 0.672$	0.736	0.204	0.704	0.111	0.678	0.075
		$\kappa_1 = 0.600$	0.618	0.176	0.609	0.092	0.599	0.063
3	58	$\phi_2 = 0.335$	0.386	0.211	0.342	0.116	0.332	0.078
		$\phi_3 = 0.672$	0.724	0.227	0.693	0.117	0.675	0.065
		$\kappa_1 = 0.400$	0.377	0.183	0.386	0.086	0.403	0.068
		$\kappa_2 = 0.200$	0.204	0.179	0.201	0.083	0.201	0.055

timates in the figure show the ordering constraint $\phi_2 < \phi_3$, which restricts the estimates to the upper left triangle of the plot. This sequence of plots shows that the estimation process consistently returned MLEs for the score parameters $\{\phi_k\}$ close to their true values (the diamond point in each plot) with reducing standard error as the sample size increases. Figures C.1 and C.2 in Appendix C.1 show similar results for the column clustering model with $C = 2$ clusters and the biclustering model with $R = 2$ and $C = 2$ clusters respectively. However, the column clustering model has the drawback that the number of $\{\alpha_i\}$ parameters is large when the sample size n is increased (e.g. 156 parameters with $n = 50$, $q = 4$ and $C = 3$) and therefore estimates would be poor with large sample sizes in that case. The consequences of this are that the standard errors are slightly higher than for row clustering and biclustering even as the number of columns m increases.

In addition, we have observed that sometimes the EM algorithm converges to

Table 4.3: Simulation study. Estimated score, row and column membership parameters for stereotype model including biclustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_c})$. The number of categories is $q = 4$. MLEs and their standard errors from the score, row and column membership parameters ($\{\phi_k\}$,$\{\pi_r\}$,$\{\kappa_c\}$) for different number of row and column clusters R and C and sample sizes n are shown.

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

a point far away from the true value. We do not notice this problem in the row clustering and biclustering versions but we detected it in approximately 5% of cases with column clustering when the sample size is $n = 50$. This problem is apparently caused by the large number of individual row parameters $\{\alpha_i\}$ in column clustering and the failure of our random starts to allow the true maximum to be found.

Our initial results described above are encouraging in their ability to estimate parameters correctly. However, we were interested to test the success of the estimation in challenging situations where it might be expected that estimation might be difficult. We chose two particular scenarios. The first case is when two of the score parameters $\{\phi_k\}$ have equal values and, therefore, from the point of view of detecting clustering, we could merge their corresponding response categories. A second scenario is to set a very small *a priori* membership probability, e.g. $\pi_2 = 0.015$, and, consequently, few data units will be classified in the related

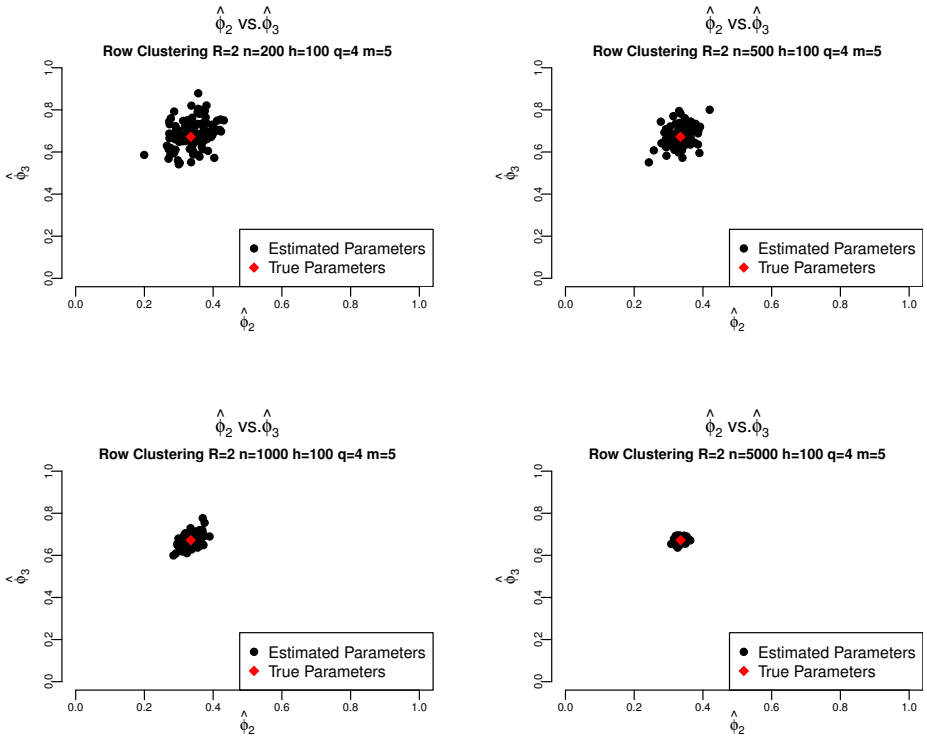

Figure 4.1: ***Simulation study***: *Convergence of $\widehat{\phi}_2$ and $\widehat{\phi}_3$ for the stereotype model including row clustering $(\alpha_r + \beta_j)$ with $R = 2$ row clusters. n, h, q, m describe the sample size, the number of replicates, the number of categories and the number of covariates respectively. The diamond point represents the true value of the parameter.*

cluster. The chosen probability must not be related to the first or last response categories because there is a relationship with the score parameters (see eq. (2.20)) and their corresponding score parameters are set to $\phi_1 = 0$ and $\phi_q = 1$ to ensure identifiability. Therefore, it is more interesting to test a free score parameter.

We have simulated these two specific scenarios for the row clustering, column clustering and biclustering models and Tables C.3-C.5 in Appendix C.1 summarises the simulation results. These are very satisfactory because our approach can identify these particular scenarios and get back values close to the true score parameters $\{\phi_k\}$ in the suite of models tested. However, some of the approximate 95% confidence intervals for the *a priori* membership probabilities $\{\pi_r\}$ do not cover their true values when the sample size is higher than $n = 1000$ and, therefore, the variability is reduced (e.g. row clustering model with $R = 4$ clusters with statistical theory (central limit theorem) providing an approximate 95% CI for π_3 and $n = 5000$ (Table C.3) is (0.262,0.298) when the true value is 0.23).

It happened less than 5% of cases, which is what we expected to happen at random. In addition, we have observed the same drawbacks described above in the column clustering version.

4.2 Real-Life Data Examples

4.2.1 Example 1: Applied Statistics Course Feedback Forms

The example is a data set with the responses of 70 students giving feedback about a second year Applied Statistics course at Victoria University of Wellington. The responses were collected in feedback forms through 10 questions (e.g. "The way this course was organised has helped me to learn"), where each question had three possible ordinal response categories: "disagree" (coded as 1), "neither agree or disagree" (coded as 2) and "agree" (coded as 3). Each question was written so that "agree" indicates a positive view of the course. The list of questions and data set are given in Tables C.6 and C.7 in Appendix C.2.

In that way, the dimensions of the data matrix Y with the responses are $n = 70$ rows (students) and $m = 10$ columns (questions) where each observation can take one of the three possible categories. Therefore, we can represent the data in a matrix as shown in Figure 4.2.

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

Figure 4.2: ***Applied Statistics course feedback forms data set:*** *The dotted circle indicates the student number 3 answered the question number 2 as "agree" (coded as 3).*

The main goal is to select the model which best represents the data, including determining the number of different groups in the data. We have fitted a suite of models from the null model (no clustering) to the main effects model and their

versions including row clustering, column clustering and biclustering. For each model, the information criteria AIC, AIC_c, BIC and ICL-BIC were computed and the results are summarised in Table 4.4.

AIC and AIC_c indicate that the best models are models with main effects ($\mu_k + \phi_k(\alpha_i + \beta_j)$) with AIC=965.26 and AIC_c=987.32, and the stereotype model version including row clustering with $R = 2$, 3 or 4 row groups ($\mu_k + \phi_k(\alpha_r + \beta_j)$). Although the main effects model is found to be the best model, for demonstration purposes we discuss here the row clustered models, which have greater interpretability. Figures 4.3-4.5 show three histograms depicting a newly-defined average of the fitted scores of student responses over the 10 questions where each student is allocated to the row group to which she/he belongs with highest posterior probability. Different shade bars represent the row cluster to which the student is assigned according to the corresponding model. This average score (along the *x*-axis) is calculated in the following way. First, we compute the fitted response probabilities with the estimated parameters over the R row clusters and the q response categories,

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

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

$$\begin{aligned} &\overline{\phi}_{rj} = \sum_{k=1}^{q} \widehat{\phi}_k P[y_{ij} = k \mid i \in r], \\ &i = 1, \ldots, n, \qquad j = 1, \ldots, m, \qquad r = 1, \ldots, R. \end{aligned} \tag{4.1}$$

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

$$\overline{\phi}_{(ij)} = \sum_{r=1}^{R} \widehat{z}_{ir} \overline{\phi}_{rj}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, m. \tag{4.2}$$

This is a numerical measure of the typical response to question j for members of row group r, appropriately adjusting for the uneven spacing of the levels of

Table 4.4: Suite of models fitted for Applied Statistics course feedback forms data set. For each information criterion, the best model in each group (no clustering, row clustering, column clustering and biclustering) is shown in boldface.

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

the ordinal response. Finally, we determine the mean of the previous weighted averages over the m columns in order to get the average fitted scores of individual i across all of the questions

$$\overline{\phi}_{(i\bullet)} = \frac{1}{m}\sum_{j=1}^{m}\overline{\phi}_{(ij)}, \qquad i = 1, \ldots, n. \tag{4.3}$$

Note that the average fitted scores of question j across all of the individuals is formulated equivalently as

$$\overline{\phi}_{(\bullet j)} = \frac{1}{n}\sum_{i=1}^{n}\overline{\phi}_{(ij)}, \qquad j = 1, \ldots, m. \tag{4.4}$$

Figures 4.3-4.5 display these $\overline{\phi}_{(i\bullet)}$ values for R =2, 3 and 4 clusters. Figures 4.3-4.4 respectively show two and three clearly distinguished groups. The histogram from Figure 4.3 presents two modes and Figure 4.4 shows two clear modes and one small mode located in the right-tale. However, Figure 4.5 where four groups are fitted shows that the fourth group only includes two students and they are not clearly distinguished from the other three groups. These graphs illustrate the conclusion from AIC/AIC$_c$ that among the row clustering models, the model with three student groups is the best for our data.

Figures 4.6 and 4.7 display the estimated probability $\widehat{\theta}_{rk}$ of a member of group r responding at category level k (eq. (2.7)). We might conclude that the students classified in the first group correspond to those with lowest opinion regarding the course, the ones in the second group have a more moderate opinion about the course and the students in the third group are those with more positive (though still heterogeneous) set of opinions.

4.2.2 Example 2: Tree Presences in Great Smoky Mountains

We use a real data set from community ecology as a second example to illustrate our likelihood-based clustering method. The data set is regarding the distribution of 41 different tree species along 12 different site stations located at altitudes between 3500 and 4500 ft and sorted by moisture level (wetter to drier). The observations consist of percentage of total tree species present at each station and was presented in R.H. Whittaker's study of vegetation of the Great Smoky Mountains

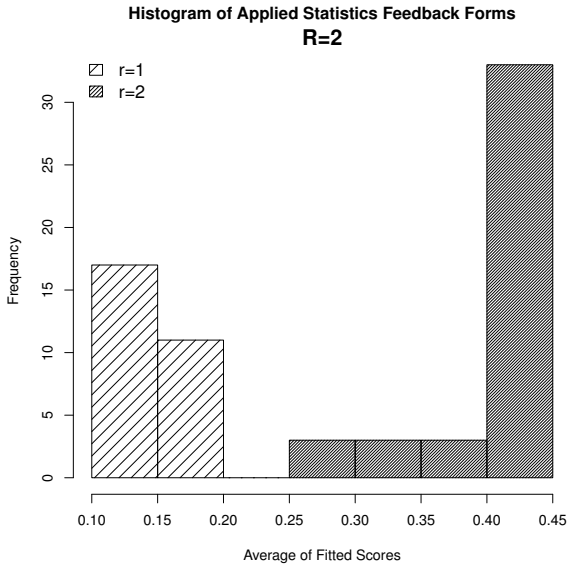

Figure 4.3: ***Applied Statistics course feedback forms data set:*** *Histogram of the $R = 2$ fitted student clusters $\{\overline{\phi}_{(i.)}\}$ from the row clustering version model.*

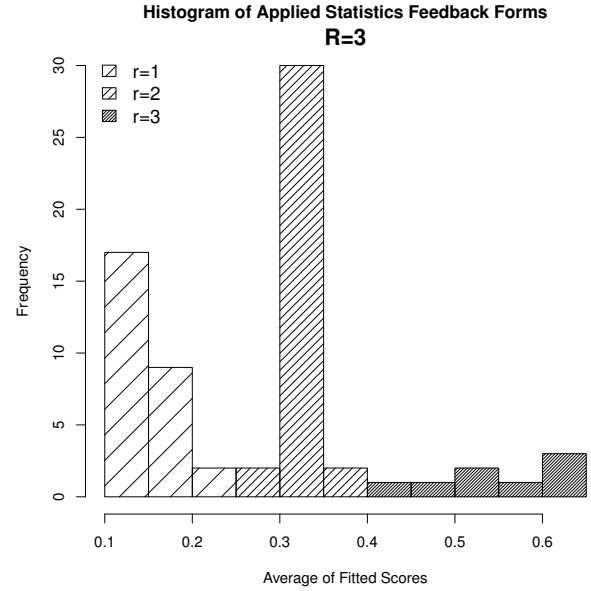

Figure 4.4: ***Applied Statistics course feedback forms data set:*** *Histogram of the $R = 3$ fitted student clusters $\{\overline{\phi}_{(i.)}\}$ from the row clustering version model.*

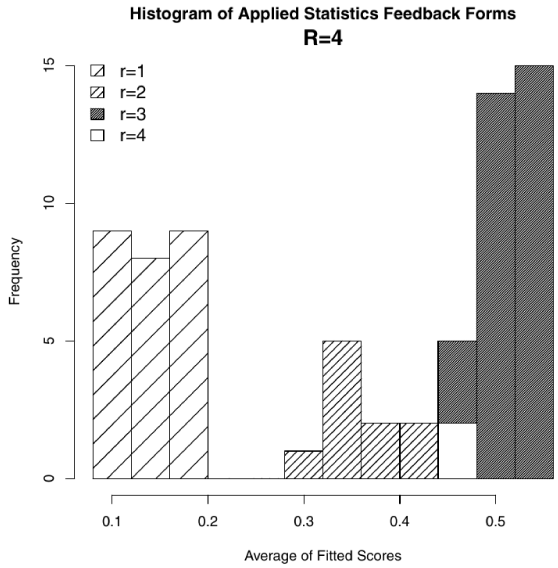

Figure 4.5: ***Applied Statistics course feedback forms data set:*** *Histogram of the $R = 4$ fitted student clusters $\{\overline{\phi}_{(i.)}\}$ from the row clustering version model.*

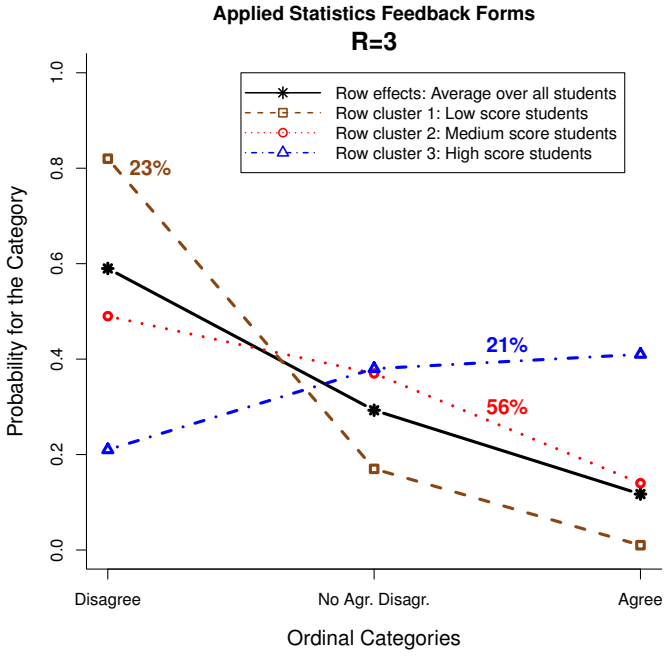

Figure 4.6: ***Applied Statistics course feedback forms data set:*** *$R = 3$ student groups. The lines depict the probability for the category $\widehat{\theta}_{r_{ijk}} = P[y_{ij} = k \mid i \in r]$ (see eq. (2.7)) for each group r and the average over all students (black line). The percentage labeling is the estimated posteriori probability $\widehat{\pi}_r$ that a student is member of each row group r (eq. (2.20)).*

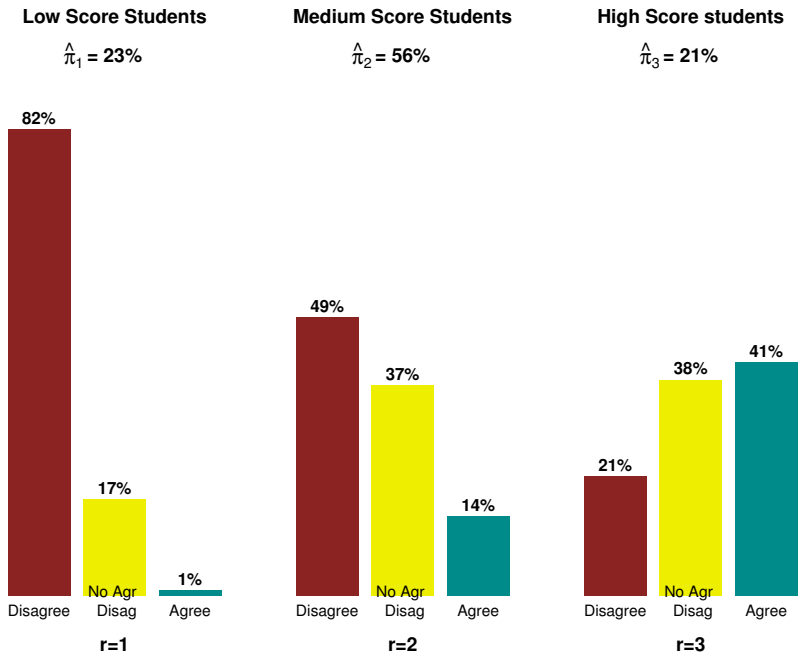

Figure 4.7: ***Applied Statistics course feedback forms data set:*** *$R = 3$ student group profiles. The percentage represents the probability $\widehat{\theta}_{r_{ijk}}$ in each category (eq. (2.7)).*

(Whittaker, 1956, Table 3). The data set is reproduced in Table C.8 in Appendix C.3.

The data include cells with a low but nonzero detection, at levels $< 0.5\%$. These missing data mean we do not have true numerical data, but only ordered data. Thus, transformations may be required (Hennig and Liao, 2013) and that presents an appropriate opportunity to replace numerical data with an ordinal scale. Section 6.2 describes advantages of using ordinal data instead of count data. In order to apply our model approach, we transform the original data $\{x_{ij}\}$ regarding tree presence percentage to ordinal response categories setting

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

based on an equitable frequency percentage for each category. Table 4.5 summarises the frequencies of tree presence data for this new ordinal scale with 5

categories. Apart from the first category, which is for sites and tree species without presences recorded, the categories with the highest frequencies are 2 and 3 (tree presence percentages between 0.5% and 8%).

Table 4.5: Frequencies of tree presence percentage by station number, in ordinal scale.

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

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

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

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

After fitting a complete set of models and comparing them by using information criteria (see the summarised results in Table C.9 in Appendix C.3), the selected model (either using AIC or AIC_c) was the stereotype model version

including row clustering with $R = 3$ row groups and with interaction factors $(\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj}))$. Figures 4.8 and 4.9 show the profiles for the three resultant row clusters. For instance, Figure 4.8 depicts that the highest probability for the row cluster number 3 (showed with a line with diamond symbols) is in ordinal category 3 and Figure 4.9 shows a first set of bars where the highest probabilities are in the categories with tree presence below 1%. Therefore, tree species classified in the first row group are those with a lower level of presence.

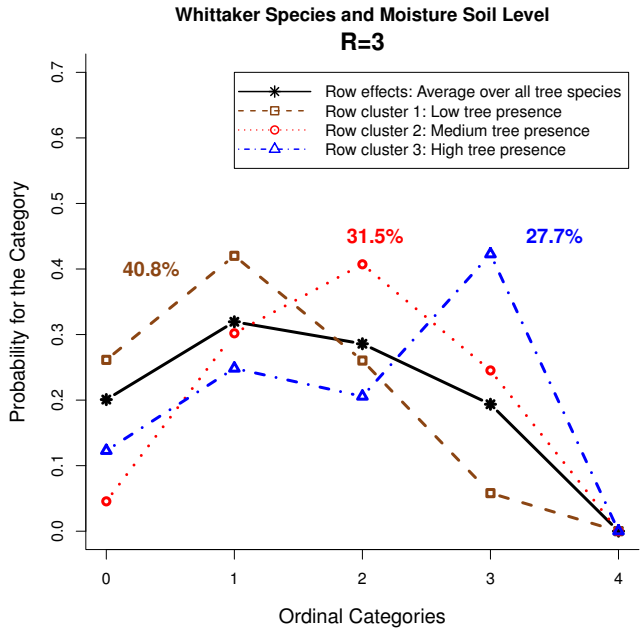

Figure 4.8: ***Tree presences data set**: $R = 3$ tree presence groups. The lines depict the probability for the category $\widehat{\theta}_{rijk} = P[y_{ij} = k \mid i \in r]$ (see eq. (2.7)) for each group r and the average over all trees (black line). The percentage labeling is the estimated posteriori probability $\widehat{\pi}_r$ that a tree species is member of each group r (eq. (2.20)).*

4.2.3 Example 3: Spider Data

The spider abundance data set (Van der Aart and Smeenk-Enserink, 1974) shows the distribution of 12 different spider species across 28 different sites. The original count data is given in Table C.10 in Appendix C.4.

The original data was categorised in order to apply the stereotype model. The

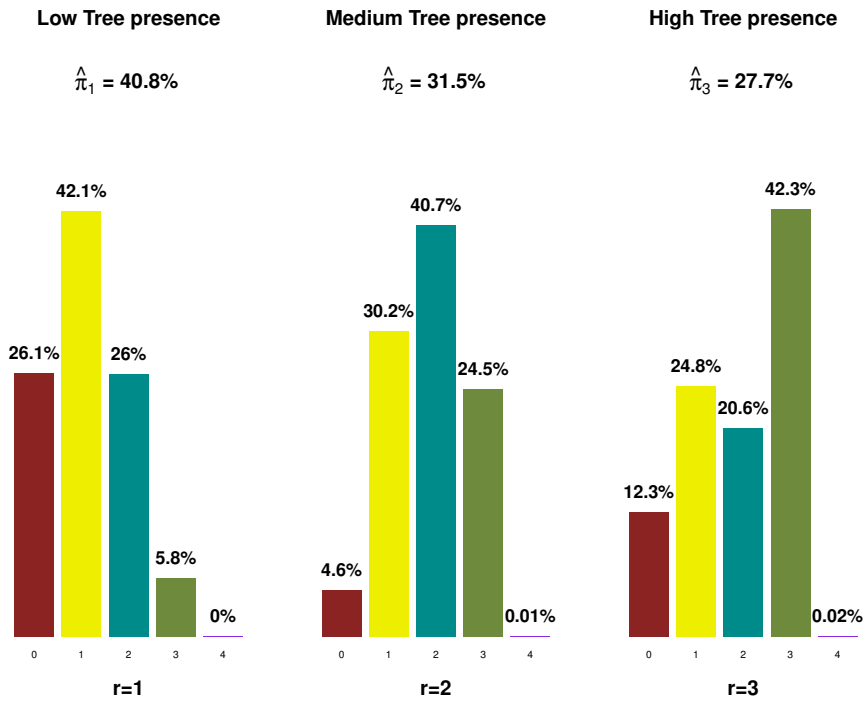

Figure 4.9: ***Tree presences data set***: $R = 3$ *tree presence profiles. The percentages represent the probability* $\widehat{\theta}_{rk}$ *in each category (eq. (2.7)).*

data was classified into 4 ordinal responses, setting:

$$y_{ij} = \begin{cases} (0) \textbf{ None} & \text{No data recorded} \\ (1) \textbf{ Low} & \text{Species coverage is below } 25\% \\ (2) \textbf{ Medium} & \text{Species coverage is between } 25\% - 65\% \\ (3) \textbf{ High} & \text{Species coverage is higher than } 65\%. \end{cases} \tag{4.5}$$

The whole ordinal data set is shown in Table C.11 in Appendix C.4. Table 4.6 summarises the frequencies of spider abundance data for this new ordinal scale. All the categories have similar frequency (between 56 and 66 observations) apart from the first category, which is for sites and spider species without presence recorded.

Table 4.6: Frequencies of spider abundance by site, in 4-level ordinal scale.

Ordinal scale	**0**	**1**	**2**	**3**	**Total**
Spider abundance	No data recorded	Low	Medium	High	Total
Frequency (y_{ij})	154	66	56	60	336

As in the previous two examples, a suite of models was fitted and information criteria measures were computed. The results are summarised in Table C.12 in Appendix C.4. Furthermore, a summary of the AIC results are in the bar plot depicted in Figure 4.11. This bar plot is sorted by AIC and the model version is distinguished by different bar colours. AIC indicates that the best model is the stereotype model version including column (sites) clustering with $C = 3$ column groups (i.e. $\mu_k + \phi_k(\alpha_i + \beta_c)$, which is labeled as $\{rn + cC3\}$ in the bar plot) with AIC$= 397.28$. Each column is allocated to the group to which the site belongs with highest posterior probability. The resultant column clustering setting is $\mathrm{C1} = \{1{-}7, 13, 14\}$, $\mathrm{C2} = \{8, 21{-}24, 27, 28\}$, and $\mathrm{C3} = \{9{-}12, 15{-}20, 25, 26\}$. Moreover, other possible models are the column clustering model with $C = 2, 4, 5$ (labeled as $\{rn + cC2\}$, $\{rn + cC4\}$, and $\{rn + cC5\}$ respectively) and the row (spider species) clustering version with $R = 2$ row clusters and an interaction factor (i.e. $\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj})$, which is labeled as $\{rR2, cm\}$).

Since each ordinal response category k $(k = 0, \ldots, 3)$ is associated with a score parameter ϕ_k, the spacing between adjacent ϕ_k values shows us how similar or different categories are (see Section 2.1.2). For this data set, the fitted score parameters were $\widehat{\phi}_0 = 0$, $\widehat{\phi}_1 = 0.39$, $\widehat{\phi}_2 = 0.89$ and $\widehat{\phi}_3 = 1$ (the end points being fixed at 0 and 1). Therefore, the distance between ordinal categories "Low" and "Medium" (0.50) is greater than that between categories "None" and "Low" (0.11) or categories "Medium" and "High" (0.39).

The scatter plot and histogram (Figure 4.10) display the average fitted scores $\{\overline{\phi}_{(.j)}\}$ over the 28 sites (see eq. (4.4)). Different colour and shape points and colour bars represent the resultant $C = 3$ column clustering setting. Three groups are distinguished in the scatter plot and the histogram presents three clear modes.

4.3 Discussion

The set of data graphical tools presented in this chapter allows us to visualise easily the different group profiles from the results of fitting likelihood-based multivariate methods for data reduction of a matrix containing ordinal data. This is a naive representation of the clustering which allows us to assess our approach in a preliminary way. Further development of data visualisation methods for finite

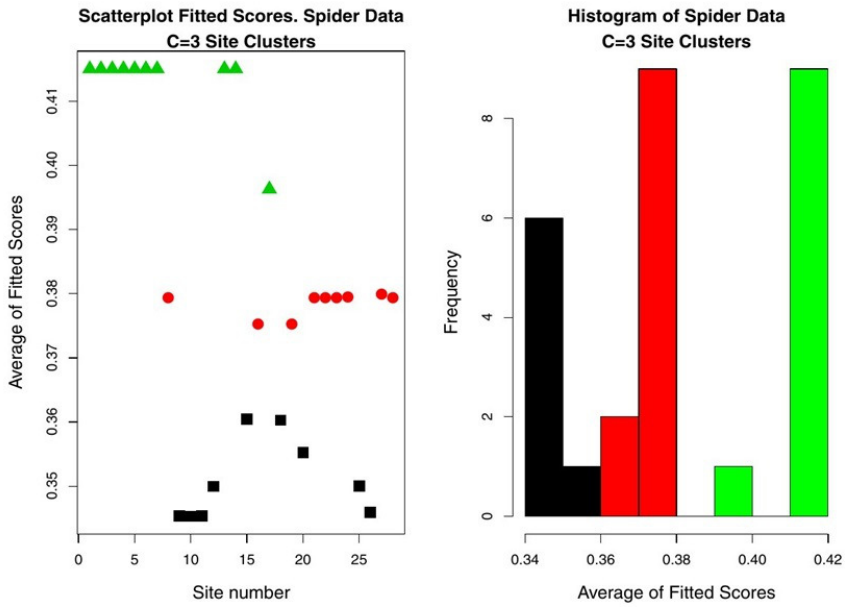

Figure 4.10: ***Spider abundance data set**: Scatter plot and histogram of the $C = 3$ fitted sites clusters $\{\overline{\phi}_{(\cdot j)}\}$ from the column clustering version model $(\mu_k + \phi_k(\alpha_i + \beta_c))$.*

mixture models based on the stereotype model is given in Chapter 5. Those new graphical tools lead to more informative visualisation. In particular, they depict the fuzzy probabilistic clustering due to the use of finite mixtures and display the possibly unequal fitted spacing among levels of the ordinal response variable.

In Chapter 6, we use the spider abundance data set from Section 4.2.3 to compare clustering results when a data set is analysed as count data in comparison to when the same data set is categorised on an ordinal scale (Section 6.4.2).

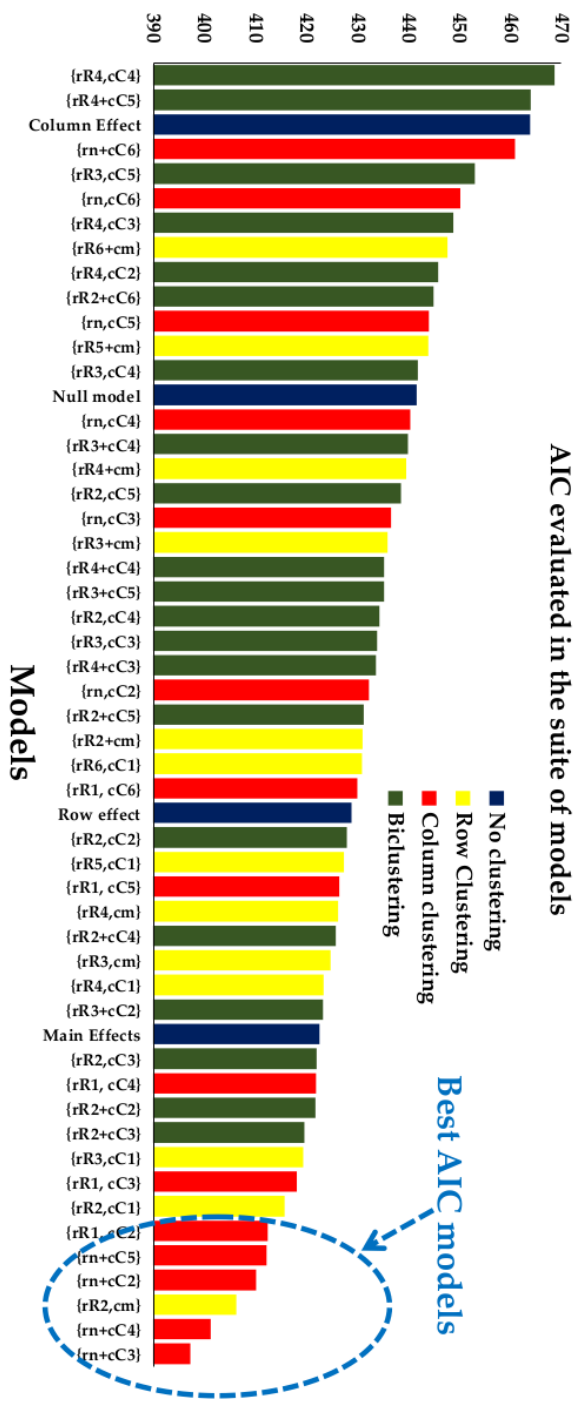

Figure 4.11: ***Spider abundance data set***: *Summary of the AIC results for the best models fitted for spider data set. The bar plot is sorted by AIC. The model version is distinguished by different bar colours: dark green bars depict biclustering models, red bars are for column clustering models, yellow bars are for row clustering models and dark blue bars represent models without clustering structure. The x-axis indicates the model which is coded as (i) homogeneous (forming a single group, denoted by rR1 or cC1), (ii) all different (denoted by rn or cp), or (iii) coming from finite mixture groups with membership unknown (rR or cC). Models with + in the labels are additive, with no interaction term.*

Chapter 5

Visualisation Techniques for Our Approach

5.1 Introduction

Using data graphical techniques allows us to present the dimensional reduction results visually and understand them more easily and quickly. Moreover, visualisation of the results is fundamental for proper communication and to improve the interdisciplinary work between statistics and other fields, making their interpretation easier. A first approach is to use exploratory data analysis (EDA) techniques to analysing data sets, e.g. summarising their main characteristics, assessing assumptions before applying statistical inference, and supporting the selection of appropriate statistical techniques. There are a number of visualisation techniques available to summarise multivariate data in fewer dimensions and to show the main features of the data, such as principal component analysis, multidimensional scaling, association analysis, correspondence analysis and ordination (or gradient analysis) (see e.g. Manly (2005); Quinn and Keough (2002)). Data visualisation is also understood as a type of unsupervised learning (see e.g. James et al. (2014, Chapter 10)). In a machine learning context, the visualisation of high dimension data sets is obtained by means of *feature extraction* techniques such as nonlinear dimensionality reduction (see e.g. Lee and Verleysen (2007); Wismüller et al. (2010)). For matrices of binary or count data, Pledger and Arnold (2014) provided likelihood-based analogues of various techniques in multivariate analysis, including multidimensional scaling, association analy-

sis, ordination, correspondence analysis, and the construction of mixture-based biplots.

In Section 4.2, we used some standard graphs such as histograms and scatter, line and bar plots to illustrate the results of the model fitting using our clustering approach, i.e. likelihood-based multivariate methods for data reduction of a matrix containing ordinal data. In this chapter, we present a set of graphs that help us to easily visualise these results. In particular, these graphs allow the visualisation of both the fuzziness in the clustering results due to the use of finite mixtures and also the spacing among ordinal categories based on the fitted score parameters $\{\widehat{\phi}_k\}$. In Section 5.2, a new graphical tool for ordinal data based on mosaic plots is introduced: the *spaced mosaic plot*. A set of visualisation tools is given in Section 5.3: graphs comparing the default equal spacing among ordinal categories and the fitted spacing based on the data (Section 5.3.1), level plots to depict a data set using the original responses (Section 5.3.2) and the fitted score parameters (Section 5.3.3), multidimensional scaling plots and ordination plots (Section 5.3.4), and contour and level plots to depict the fuzziness of the results (Section 5.3.5).

5.2 Spaced Mosaic Plots

5.2.1 Mosaic Plot. Description

The mosaic plot was developed by Hartigan and Kleiner (1981) and refined by Friendly (1991). It is a graphical method for visualising data from two qualitative variables which gives an overview of the data, makes it possible to recognize relationships and show the cross-sectional distribution of different variables. In our case, we consider the ordinal response variable and the number of fitted clusters in the data as those two qualitative variables. For instance, an ordinal data matrix following a four-category Likert scale ("Disagree","No Opinion","Agree","Strongly Agree") clustered into three row clusters is depicted as a mosaic plot in Figure 5.1. The mosaic plot is divided into 3 horizontal bands over the y-axis (one for each row cluster) and 4 vertical bands over the x-axis (one for each ordinal response category). The areas represent the frequencies as explained in Section 5.2.2.

One improvement we can incorporate in a mosaic plot due to the use of the

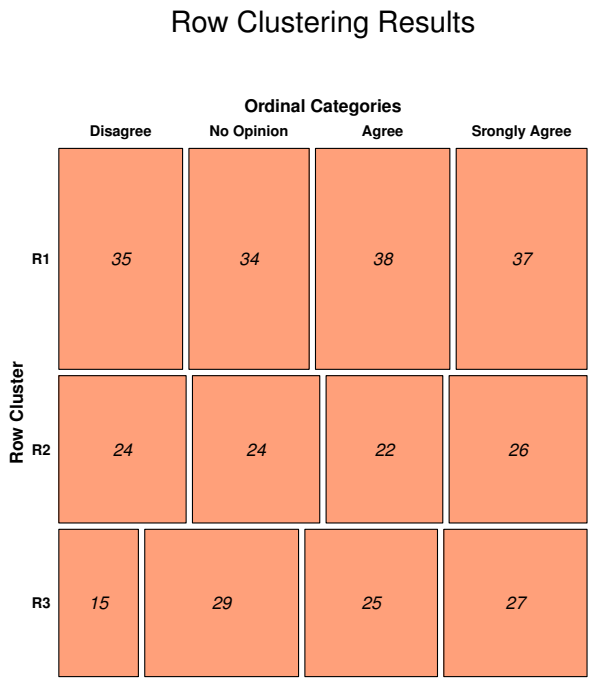

Figure 5.1: ***Mosaic plot:*** *Plot including row cluster structure* $R = 3$ *and* 4 *ordinal categories.*

ordinal stereotype model is the estimation of score parameters $\{\phi_k\}$. Those parameters determine the space between two adjacent ordinal categories based on the data (see Anderson (1984); Agresti (2010) for more detail). For instance, the space between "Disagree" and "No Opinion" can be higher than the space between "Agree" and "Strongly Agree". The inclusion of space within a regular mosaic plot generates an improved graph with more information which we called the spaced mosaic plot and which is developed in the following two sections.

5.2.2 Spaced Mosaic Plot. Description

We use an ordinal real data set from community ecology as an example to illustrate the spaced mosaic plots in the case of clustering the rows. The data set is regarding the distribution of 77 different angiosperms along 30 different sites. The study was carried out at Bola Heights in Royal National Park, about 37 km south-west of Sydney and 200 meters above sea level (see Tozer and Bradstock (2002) for more detail). The goal of this vegetation survey data is to group species

observations to derive community types. The 2310 ordinal observations consist of the level of angiosperm species (row) presence at each site (column) in combination with the percentage of coverage within the site. Thus, the ordinal scale in this data follows a cover/abundance estimate using a modified Braun-Blanquet scale (Westhoff and van der Maarel, 1978) determined by the field worker:

$$\begin{cases} 0 & \text{no data recorded} \\ 1 & \text{one/a few individuals and less than 5\% cover} \\ 2 & \text{uncommon and less than 5\% cover} \\ 3 & \text{common/very abundant and less than 5\% cover} \\ & \text{or coverage higher than 5\%.} \end{cases}$$

After fitting a complete set of models and comparing them by using the Akaike information criteria (AIC, Akaike (1973)), the selected model was the stereotype model version including row (angiosperm species) clustering with $R = 4$ species groups. Each species i is assigned uniquely to a row group with the highest posterior probability $\widehat{Z}_{ir}$. Figures 5.2-5.4 show the results for this example. Firstly, Figure 5.2 depicts the raw data without including row clustering, Figure 5.3 depicts the data including row cluster structure and Figure 5.4 depicts the data including both row cluster structure and fitted spacing between ordinal categories. A comprehensive description for each Figure is as follows:

- Figure 5.2 shows the overall distribution of ordinal responses over all the cells, ignoring rows and columns. Thus, area is simply equivalent to frequency of each response value across the whole dataset. The ordinal category 0 response is most common by far, and ordinal category 3 the least.

- Figure 5.3 shows the clustering in the rows, putting each species into one of four row clusters according to the distribution of ordinal responses across the columns (sites) of the original data matrix. This divides the plot into four horizontal bands, one for each row group. The height of each band is proportional to the number of rows in the group. Therefore, we can see that row groups 1 and 4 ($R1$ and $R4$) are the largest, much larger than row groups 2 and 3 ($R2$ and $R3$). Within each row group we represent the frequencies of the four ordinal responses by the area of each block. Angiosperm species

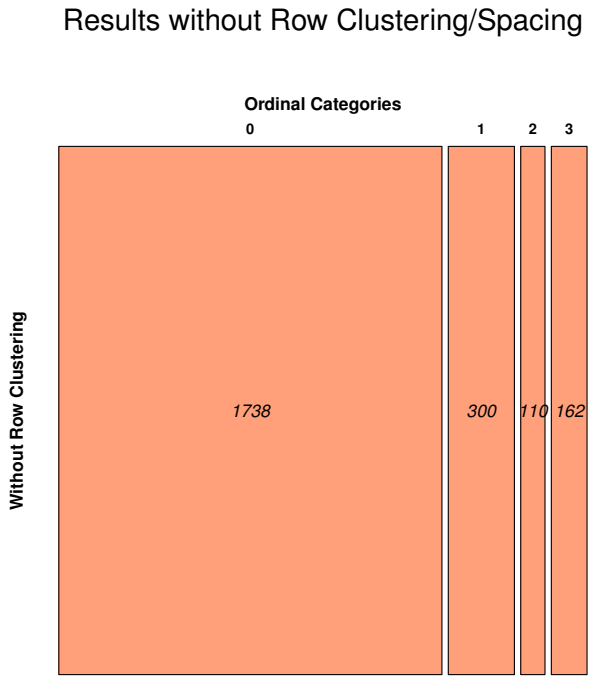

Figure 5.2: ***Angiosperm data set:*** *Mosaic plot without spacing or row clustering.*

of row group 4 show a strong preference for ordinal response categories 0 and 1, and rarely respond at category 2 or 3. Contrast this with row group 2, which has 50% of its responses at ordinal category 3. Note that this diagram does not in any way show the ordering across the sites (columns) – it is simply a pooling of frequencies of all of the responses for species in the same row group.

- Figure 5.4 takes the bands and blocks from Figure 5.3, but separates them out to indicate the numerical spacing between the response categories that the model has identified. Since each ordinal response category is associated with a score parameter ϕ_k $(k = 0, \ldots, 3)$ the spacing between these ϕ_k values shows us how similar or different adjacent categories are. In this model the fitted score parameters are $\phi_0 = 0$, $\phi_1 = 0.66$, $\phi_2 = 0.96$ and $\phi_3 = 1$ (the end points being fixed at 0 and 1). The distance between category 0 and category 1 (0.66) is much greater than that between categories 1 and 2 (0.30) or categories 2 and 3 (0.04). In each row group band, we have inserted space and a different colour oblong proportional to these differences between two adjacent ordinal categories. For instance, a yellow oblong of the same width

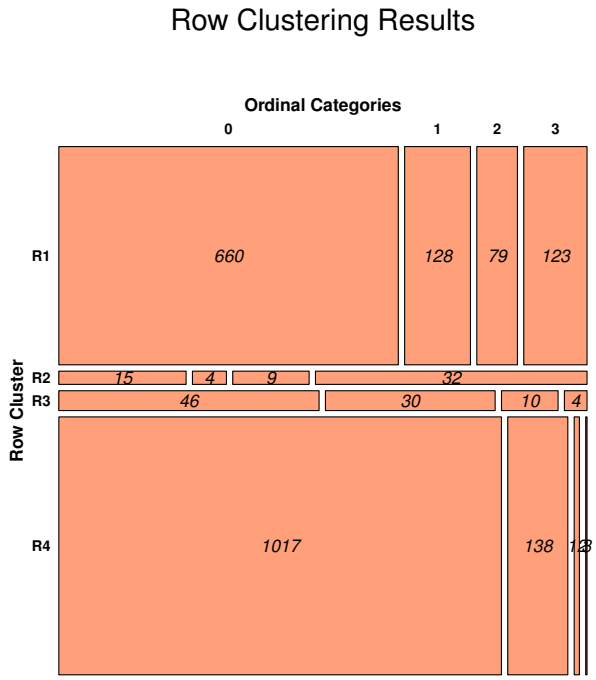

Figure 5.3: ***Angiosperm data set:*** *Mosaic plot including row cluster structure* $R = 4$.

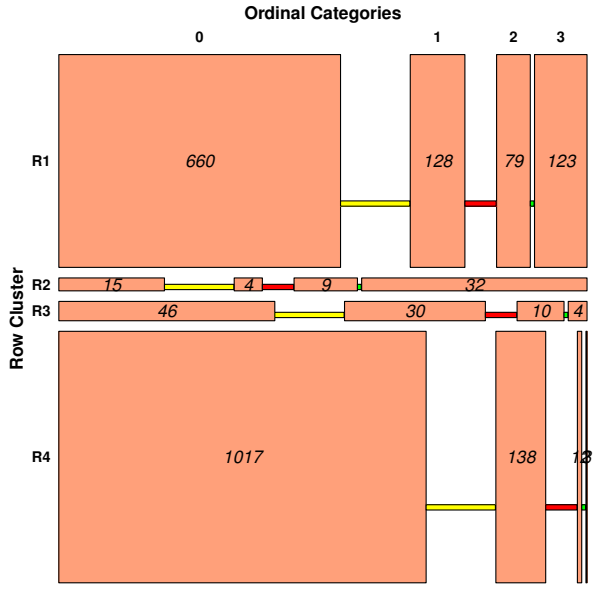

Figure 5.4: ***Angiosperm data set:*** *Mosaic plot with spacing for the row clustering model* $R = 4$.

has been inserted between categories 0 and 1 in each band. Note that these oblongs do not line up vertically with each other between bands due to the differing counts at category 0, nevertheless oblongs of the same colour are the same width. In so doing, we can immediately see that categories 2 and 3 are close to each other, without needing to refer to the numerical values of ϕ_k. Inspection of Figure 5.4 might lead us to conclude that categories 2 and 3 are so similar that these two groups might just as well be collapsed into a single group.

5.2.3 Outlining Spaced Mosaic Plots

The main features of a spaced mosaic plot are:

- spread along the x-axis represents the ordinal categories in the data and the y-axis represents the row clustering obtained by our methodology. The data frequency of each combination in terms of ordinal category and row cluster is shown by the area of each box.

- The greater the area of a specific box, the higher the proportion of data allocated to the related ordinal category. For instance, the box located on the top right depicts the proportional number of species (angiosperms) allocated in the first cluster ($R1$) and with Braun-Blanquet scale 3.

- The greater the height of a specific box, the higher the proportion of rows classified in that particular row group. For example, the bottom left box corresponding to the row cluster 4 ($R4$) and the ordinal category 0 is the widest and the highest because it contains 1017 combinations of species-samples over 2310 (44%). None of the other boxes have higher frequencies.

- The spacing between two levels of the ordinal categories (x-axis) is dictated by the data. It represents the proximity of two adjacent ordinal categories. Determining the distance among ordinal categories is a key advantage of the stereotype model in comparison with other similar methods.

Spaced mosaic plots allow us to see the at once the relative sizes of the row groups, the relative frequencies of the different response categories within each row group and the differences between the levels of the response categories.

However they do not show the fuzziness of the clustering. We discuss methods of showing this in Section 5.3.5 below.

The documentation of an **R** function we developed to generate spaced mosaic plots is presented in Appendix D. In addition, a technical report introducing these plots is published in Fernández et al. (2014b).

5.3 Other Visualisation Tools

5.3.1 Reassigned Ordinal Scale

The categories in an ordinal response variable have been labeled as $k = 1, 2, \ldots, q$ throughout this thesis. For instance, we coded the ordinal response categories "disagree", "neither agree or disagree" and "agree" as $1, 2,$ and 3 respectively in the Applied Statistics example in Section 4.2.1. The use of the first q positive integers as labels does not imply that there is equal spacing among ordinal categories. The fitted spacing is instead determined by the distance among adjacent score parameters $\{\widehat{\phi}_k\}$. Given an ordinal response variable, the purpose is to develop a visualisation tool that allow us to compare visually the default equal spacing among its categories with the fitted spacing dictated by the data.

Figure 5.5 depicts two graphs with a 5 level Likert scale in an ordinal response variable (i.e. "Strongly Disagree", "Disagree", "No opinion", "Agree", "Strongly Agree"). In the right graph, the equally spaced scale is depicted in blue axis and the fitted score scale in green axis. The fitted score parameters were $\widehat{\phi}_2 = 0.252, \widehat{\phi}_3 = 0.748,$ and $\widehat{\phi}_4 = 0.946$ ($\phi_1 = 0$ and $\phi_5 = 1$ are restricted to ensure identifiability). The left graph shows a dotted blue straight line which corresponds to the equally spaced categories and the green line depicts how different the fitted score parameters are from this uniformity. The amount of non-linearity shows the distortion of the scale from the incorrect equally-spaced scale. Therefore, the adjacent categories are not equally spaced based on the data. The spacing between categories "Disagree" and "No opinion" is the largest and the shortest is between "Strongly Agree" and "Agree" categories. Thus, these two graphs allows us to easily depict the new uneven spacing of the levels of the ordinal response, dictated by the data.

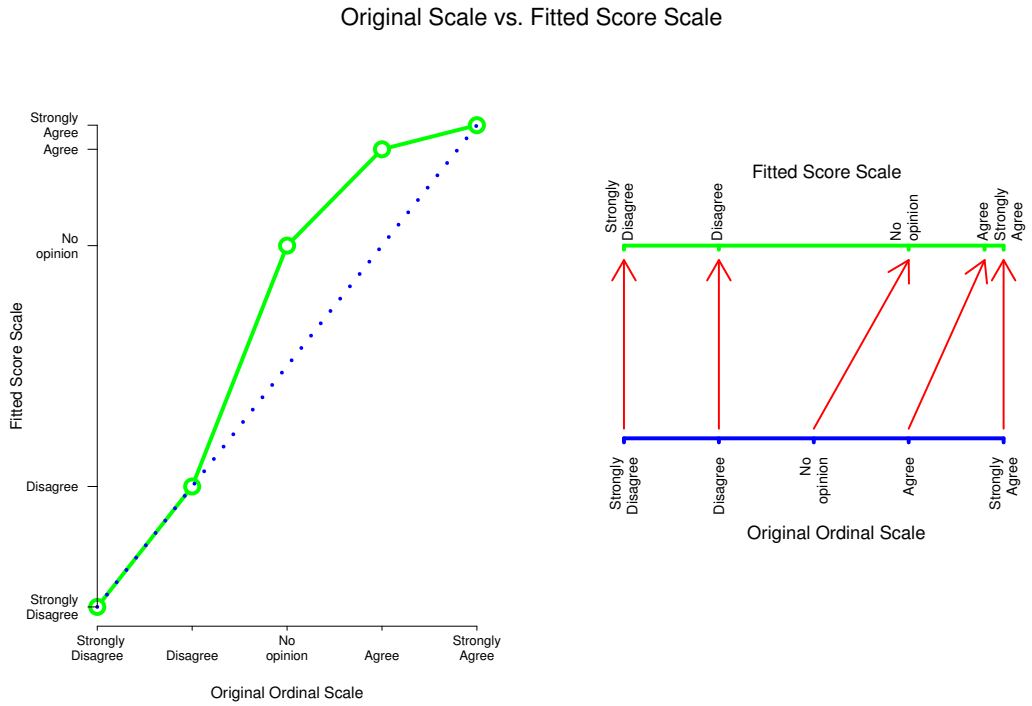

Figure 5.5: ***Reassigned ordinal scale:*** *Scale comparison between default equal spacing and fitted spacing given by score parameters* $\{\widehat{\phi}_k\}$ *for ordinal response variable with 5 level Likert scale categories ("Strongly Disagree", "Disagree", "No opinion", "Agree", "Strongly Agree").*

5.3.2 Data Set Level Plots

Patterns in the ordinal responses of the rows (e.g. subjects) and the columns (e.g. questions) may be visualised in coloured level plots, each ordinal response level being represented by colours from a chosen palette. The Applied Statistics data set described in Section 4.2.1 is used to illustrate these graphs throughout this section. This data set $Y = \{y_{ij}\}$ shows the ordinal responses of $n = 70$ students (rows) through $m = 10$ questions (columns). The number of ordinal categories is $q = 3$: "agree", "neither agree or disagree", and "disagree". We concluded that the best fitted clustering model for this data set is a row clustering with $R = 3$ student groups (see the model fitting results in Table 4.4 in Section 4.2.1). The student allocation in each group based on the highest posterior probability criterion is:

$$
\begin{aligned}
R1 =& \{8, 11, 14, 15, 18, 19, 21, 23, 24, 25, 27, 28, 31, 32, 39, 41, 42, 48, \\
& 49, 53, 55, 56, 58, 59, 61, 62, 65\}, \\
R2 =& \{1-5, 7, 9, 10, 13, 16, 17, 20, 22, 26, 29, 30, 33-38, 40, 43, 44, 46, \\
& 47, 50, 51, 52, 57, 63, 64, 66, 67-69\}, \text{and} \\
R3 =& \{6, 12, 45, 54, 60, 70\}.
\end{aligned}
\tag{5.1}
$$

Figure 5.6 shows the original data set without any row or column rearrangement. Dark green cells represent students answering the corresponding question as "disagree", the light green ones are related to the "neither agree or disagree" category, and the light brown ones to the "agree" category. It seems that the students' tendency was to respond in lower levels in questions 1-3 and 9 (mostly dark and light green cells) and more combined in the other questions. This data

Figure 5.6: ***Applied Statistics course feedback forms data set:*** *Level plot depicting the data responses $\{y_{ij}\}$ of 70 students (y-axis) through 10 questions (x-axis).*

set sorted by student (row) according to the row cluster structure given in (5.1) is presented in Figure 5.7. The blue lines across the figure divide the level plot to distinguish the clusters (y-axis). Thus, the bottom cluster corresponds to the

student group $R1$, the middle one is the group $R2$, and the top one corresponds to $R3$. The students within each cluster are also sorted by their ordinal responses $\{y_{ij}\}$ to better illustrate the smooth transition between clusters, i.e. the students with lower responses (dark green cells, y_{ij} ="disagree") within a particular cluster are allocated at the bottom of that group, and those with more moderate (light green cells, y_{ij} ="neither agree or disagree") and higher responses (light brown cells, y_{ij} ="agree") are allocated higher up. We might conclude that the students

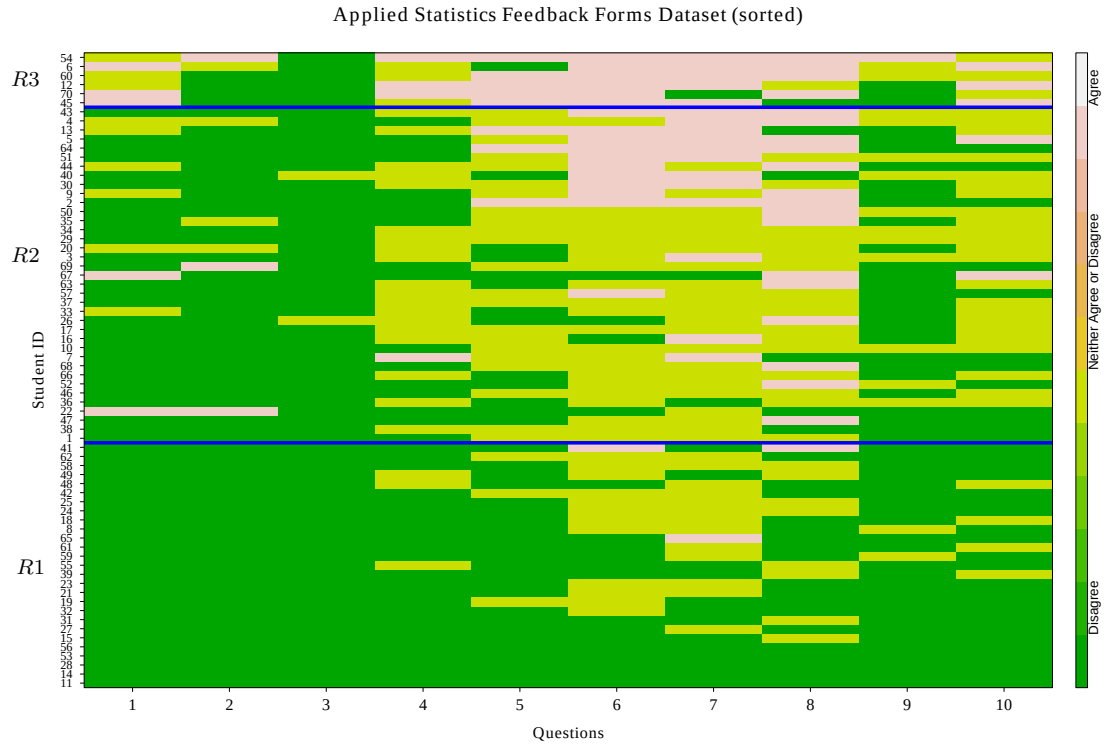

Figure 5.7: ***Applied Statistics course feedback forms data set:*** *Level plot depicting the data responses $\{y_{ij}\}$ of 70 students (y-axis) through 10 questions (x-axis), sorted by student cluster with $R = 3$ groups. The blue lines across divide the plot to show the 3 clusters ($R1$, $R2$, and $R3$). The students within each cluster are also sorted by their ordinal responses $\{y_{ij}\}$ to illustrate better the smooth transition between clusters.*

from cluster $R1$ correspond to those with lowest opinion regarding the course (mostly dark green cells), the ones in the cluster $R2$ have a more moderate opinion about the course (the colour of the cells is quite balanced between light and dark green and light brown) and the students in the group $R3$ are those with more positive (though still heterogeneous) set of opinions (more light brown cells).

5.3.3 Level Plot Based on the Score Parameters

In Section 5.3.1, we developed graphs to compare the default equal spacing among ordinal categories with the uneven fitted spacing dictated by data. This fitted spacing determined by the score parameters $\{\widehat{\phi}_k\}$ can be incorporated in the level plots introduced in the previous section.

The graph of the Applied Statistics data set in Figure 5.8 is similar to Figure 5.7. However, instead of the original ordinal scale responses $\{y_{ij}\}$, this figure shows the mean response level $\{\overline{\phi}_{(ij)}\}$ of student i ($i = 1, \ldots, n$) to question j ($j = 1, \ldots, m$), conditional on its fuzzy allocation to the row clusters (see eq. (4.2) in Section 4.2 for reviewing how this average score is calculated). This is a numer-

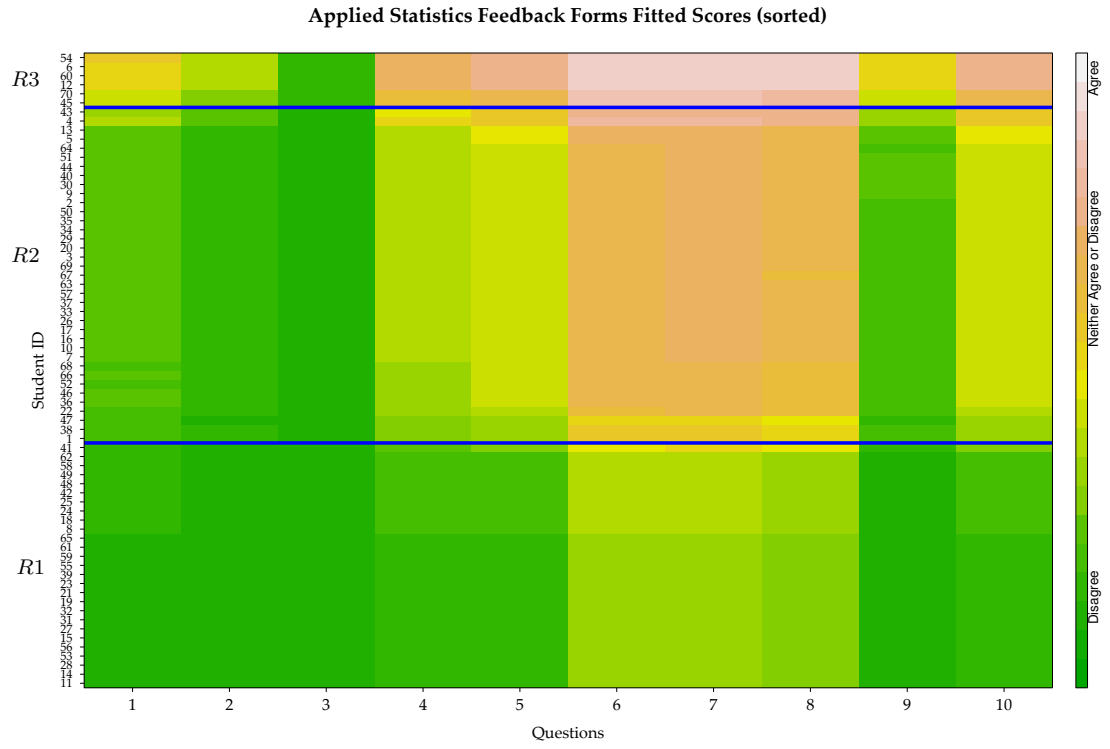

Figure 5.8: ***Applied Statistics course feedback forms data set:*** *Level plot depicting mean response level of each student to each question, conditional on its fuzzy allocation to the $R = 3$ row clusters $\{\overline{\phi}_{(ij)}\}$ (eq. (4.2)). The horizontal blue lines divide the plot to show the 3 clusters ($R1$, $R2$, and $R3$). The students within each cluster are also sorted by their ordinal responses $\{y_{ij}\}$ to illustrate better the smooth transition between clusters.*

ical measure of the typical response to question j ($j = 1, \ldots, m$) for members of student group r ($r \in \{R1, R2, R3\}$), appropriately adjusting for the uneven spac-

ing of the levels in the ordinal response variable. The finest variation cell colour in this graph is based on a terrain palette which goes from dark green to light brown. Thus, dark green cells represent students with lowest opinion regarding the course and light brown cells are those with more positive view. We have intentionally chosen this colour palette with the aim of comparing this graph to Figure 5.7. This comparison is shown in the side-by-side Figure 5.9.

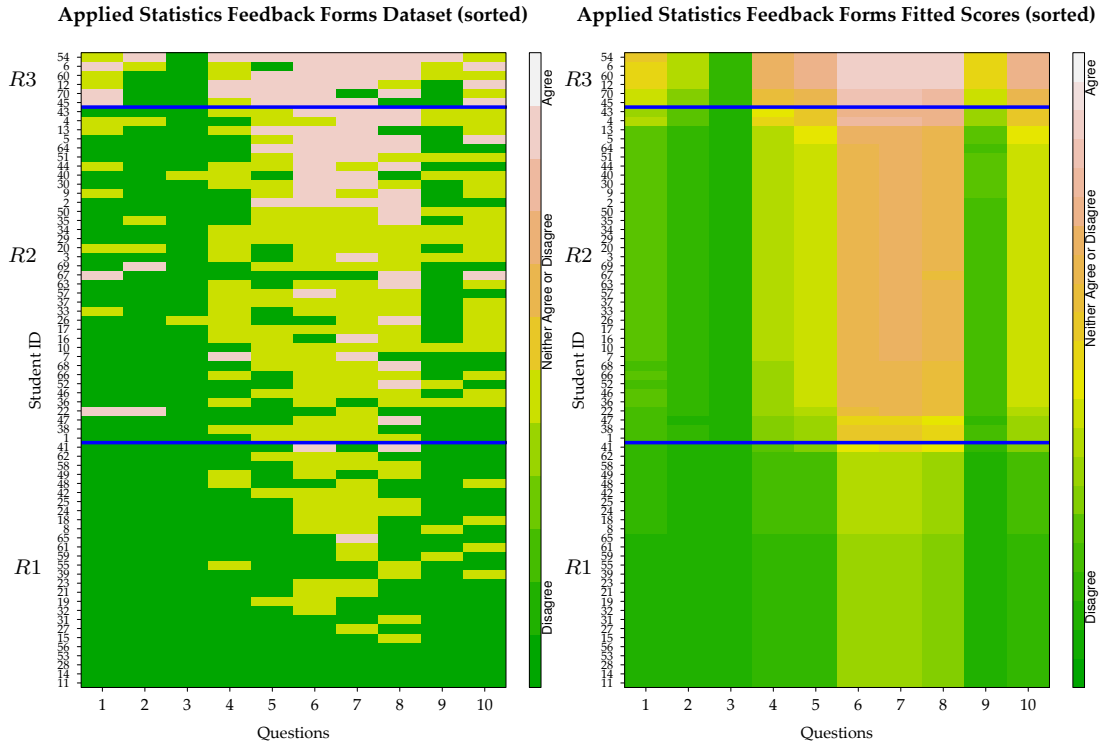

Figure 5.9: ***Applied Statistics course feedback forms data set:*** *Level plots depicting response level of each student (y-axis) to each question (x-axis). Level plot on the left presents the data responses* $\{y_{ij}\}$*. Level plot on the right presents the fitted ordinal scaled based on the weighted measure (4.2). The blue lines across divide the plot to show the 3 clusters (*$R1$*,* $R2$*, and* $R3$*). The students within each cluster are also sorted by their ordinal responses* $\{y_{ij}\}$ *to illustrate better the smooth transition between clusters.*

The left level plot presents the data set responses $\{y_{ij}\}$ and the one on the right depicts the weighted numerical measure $\{\overline{\phi}_{(ij)}\}$ based on the fitted score parameters: $\widehat{\phi}_2 = 0.252, \widehat{\phi}_3 = 0.748$, and $\widehat{\phi}_4 = 0.946$ ($\phi_1 = 0$ and $\phi_5 = 1$ are restricted to ensure identifiability). This comparison is similar to the one shown in Figure 5.5 between the original and the fitted ordinal scale but generalizing the latter scale to a measure for each student and question. The horizontal blue

lines divide both plots to show the clusters according to the row cluster structure given in (5.1). Both graphs show a similar pattern but the right level plot has a smoother appearance and a finer variation in colours because it is depicting a continuous numerical measure while the left level plot is a discrete three-level scale. Therefore, we can observe clearer differences in the right level plot. For instance, it is difficult to detect the differences in question 9 for $R2$ and $R3$ (mid and top clusters on the graph respectively) in the left graph. However, it is easier to identify them in the right graph (more light green tone in $R3$ than $R2$).

5.3.4 Multidimensional Scaling Scatter Plots

The level plots in the previous section were obtained by depicting the mean response level $\{\overline{\phi}_{(ij)}\}$ of row i ($i = 1, \ldots, n$) to column j ($j = 1, \ldots, m$). In order to calculate this numerical measure, the $R \times m$ matrix $\{\overline{\phi}_{rj}\}$ of weighted average over the q ordinal categories for each row cluster is obtained (see eq. (4.1) in Section 4.2 for the definition of this matrix). We can use any pair of rows (clusters) in this matrix to depict a 2D *multidimensional scaling plot* (MDS) of the m columns.

Figure 5.10 presents three MDS plots for all possible pairs of clusters for the Applied Statistics data set according to a fitted row clustering model with $R = 3$ student clusters ($R1$ vs. $R2$, $R1$ vs. $R3$, and $R2$ vs. $R3$). Each plot depicts $\{\overline{\phi}_{rj}\}$ and $\{\overline{\phi}_{r'j}\}$ for row groups $r, r' \in \{R1, R2, R3\}$. These plots of questions show similar patterns. We note that questions 6-8 are plotted together in all three MDS plots illustrating their similarities to each other, and differences from the other 7 questions (1-5 and 9-10). Likewise questions 5 and 10 and questions 1 and 9 coincide graphically in the three plots, illustrating that they are associated.

Each row $\{\overline{\phi}_{rj}\}$ provides a separate one-dimensional *ordination* of the questions. These are the projections for each row cluster onto the axes in Figure 5.10 and are shown individually in Figure 5.11. We note that the range of values of the ordination of the questions is different depending on the row group. Questions in the axis pertaining to row cluster $R1$ (Figure 5.11(a)) are more concentrated in lower weighted average $\{\overline{\phi}_{rj}\}$ values which shows mostly levels of disagreement in student responses. On the other hand, questions in axis $R3$ (Figure 5.11(c)) are located in higher values depicting levels of agreement and those in axis $R2$ (Figure 5.11(b)) show more moderate student answers. In addition to this figure, Figure 5.12 depicts the weighted average $\{\overline{\phi}_{rj}\}$ for the $R = 3$ row (student) clus-

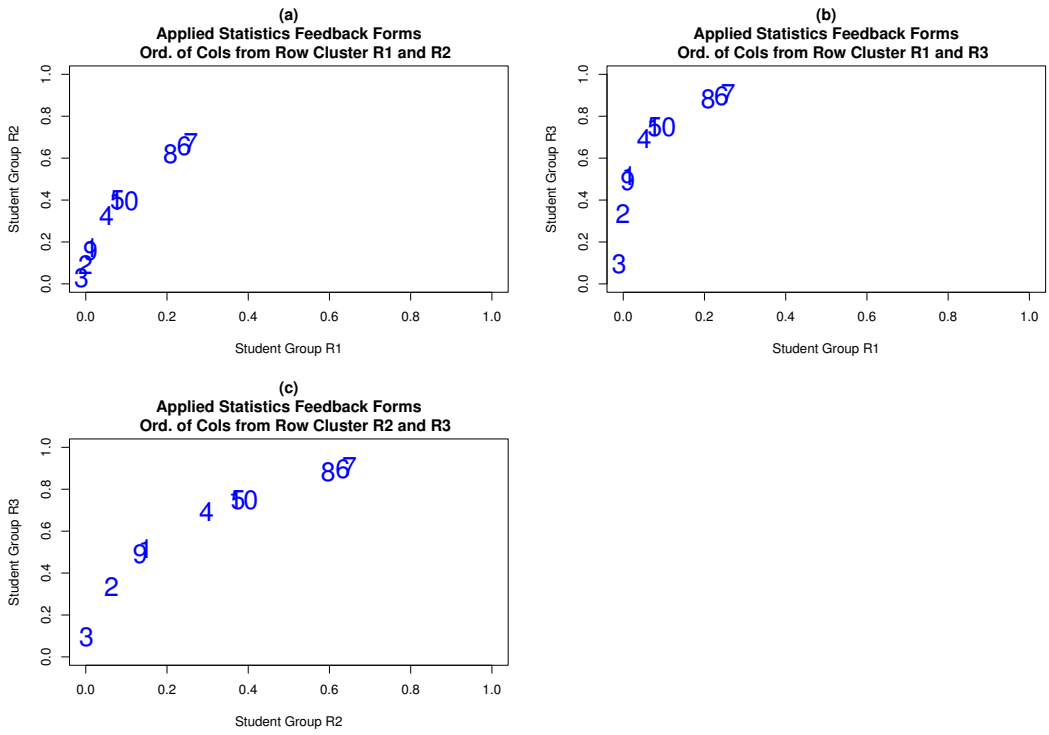

Figure 5.10: ***Applied Statistics course feedback forms data set:*** *Multidimensional scaling plots of the $m = 10$ questions using the row cluster structure given in (5.1). The axes are the weighted average $\{\overline{\phi}_{rj}\}$ (eq. (4.1) in Section 4.2) for two row groups. Each plot depicts $\{\overline{\phi}_{rj}\}$ and $\{\overline{\phi}_{r'j}\}$ for row groups $r, r' \in \{R1, R2, R3\}$. The left top plot (a) is for student clusters $R1$ and $R2$, the right top (b) is the one for $R1$ and $R3$, and the left bottom one (c) corresponds to $R2$ and $R3$.*

ters broken down by each question ($m = 1, \ldots, 10$). Each axis corresponds to a question and shows the profile of the responses for a particular student group. For instance, the students in cluster $R1$ have the lowest opinion on average regarding the course in question 5 ($\overline{\phi}_{15} = 0.09$) in comparison with students in cluster $R2$ ($\overline{\phi}_{25} = 0.39$) and $R3$ ($\overline{\phi}_{35} = 0.74$). We note that the responses are very consistent in all the questions apart from question 3: cluster $R1$ are consistently low responses, $R2$ intermediate, and $R3$ high responses. In question 3, responses from $R3$ and, particularly, $R2$ are as low as those from $R1$. We also note again that questions 6-8 are similar as clusters $R1$-$R3$ are in the same place along these axes.

From the last two figures, we can calculate an overall ordination of clusters

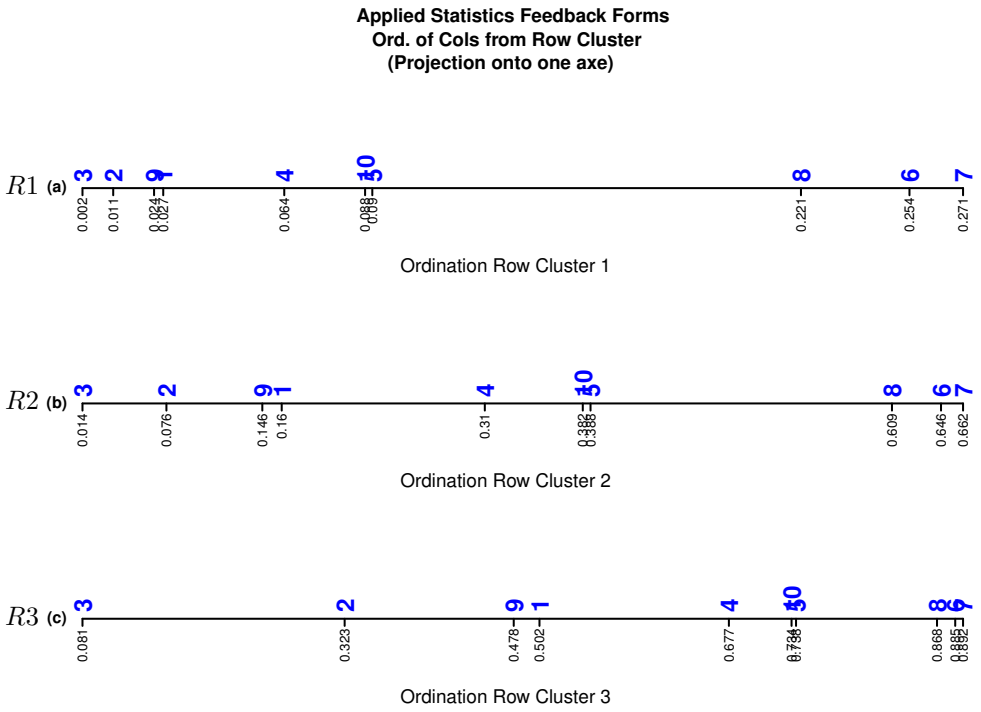

Figure 5.11: ***Applied Statistics course feedback forms data set:*** *Projections onto an axis of the $m = 10$ questions using the row cluster structure given in (5.1). Each axis is related to one row (student) cluster and depicts the ordination of the question based on the weighted average $\{\overline{\phi}_{rj}\}$ (eq. (4.1) in Section 4.2) from the MDS plots in Figure 5.10. Figure (a) is related to row cluster $R1$, Figure (b) to row cluster $R2$ and Figure (c) to row cluster $R3$. Note that each axis has different ranges.*

(from Figure 5.11):

$$\overline{\overline{\phi}}_r = \frac{1}{m}\sum_{j=1}^{m}\overline{\phi}_{rj}, \qquad r = 1, \ldots, R,$$

and also an overall ordination of the questions (from Figure 5.12):

$$\overline{\overline{\phi}}_j = \frac{1}{R}\sum_{r=1}^{R}\overline{\phi}_{rj}, \qquad j = 1, \ldots, m.$$

For instance, the overall ordination of question 1 is $\overline{\overline{\phi}}_1 = \frac{0.03+0.16+0.5}{3} = 0.23$ and of question 6 is $\overline{\overline{\phi}}_6 = \frac{0.25+0.65+0.89}{3} = 0.60$.

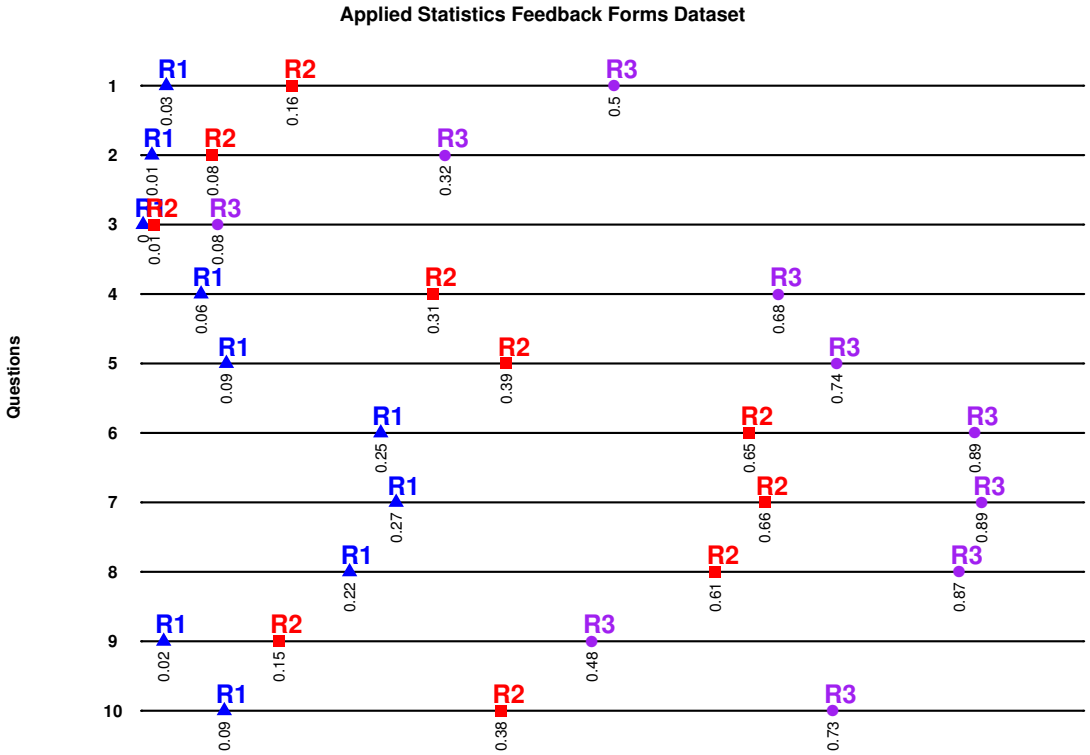

Figure 5.12: ***Applied Statistics course feedback forms data set:*** *Projections onto an axis of the weighted average* $\{\overline{\phi}_{rj}\}$ *(eq. (4.1) in Section 4.2) for each question. Each axis depicts the weighted average for the* $R = 3$ *row (student) clusters related to a question. Cluster* $R1$ *is shown in blue triangles,* $R2$ *in red squares and* $R3$ *in purple circles.*

5.3.5 Contour and Level Plots to Represent Fuzziness

Our finite mixture approach performs a fuzzy assignment of rows and/or columns to clusters based on the posterior probabilities, as we presented in its model formulation (Chapter 2). In this section, different visualisation tools to represent this fuzzy probabilistic clustering are presented. In particular, the fuzziness is depicted in 3 graphs which are based on the allocation of each row i from the data $\{y_{ij}\}$ in each cluster r, the distances among score parameters $\{\widehat{\phi}_k\}$, and the membership posterior probabilities $\{\widehat{Z}_{ir}\}$ that row i is in cluster r once we have observed the data $\{y_{ij}\}$.

Crisp Clustering Contour Plot

Once the best fitted clustering model for a data set is identified and the clustering allocation of the rows (or columns) is determined, a crisp clustering of the rows (i.e. each row allocated with probability one to its row group) can be easily displayed.

For the Applied Statistics data set, we identified $R = 3$ different row cluster profiles: a student group with the higher responses, another group with the intermediate ones and a third student group with the lower ones (see Section 4.2.1). The student allocation in each cluster is based on highest posterior probability criterion:

$$\widehat{r}_i = \underset{r \in 1,\ldots,R}{\text{argmax}} \ \widehat{Z}_{ir}, \quad i = 1, \ldots, n.$$

Given the crisp cluster structure we can calculate the distance between any pair of row cluster label allocations: $d_{ii'} = |\widehat{r}_i - \widehat{r}_{i'}|$ $(i, i' = 1, \ldots, n)$. For instance, if a particular student i is classified in the cluster $R1$ and another student i' is allocated in the cluster $R3$ then the value to depict is $d_{ii'} = |\widehat{r}_i - \widehat{r}_{i'}| = |1 - 3| = 2$. This way of calculating the distance $d_{ii'}$ is based on a consecutive numbering of the clusters (i.e. $R1$ is the first cluster, $R2$ is the second one, and so on). The graphs in this section are calculated based on this distance. However, the numbering of the clusters might be arbitrary and, in that case, a better distance would be $d_{ii'} = \left| \frac{1}{m} \sum_{j=1}^{m} (\overline{\phi}_{\widehat{r}_i j} - \overline{\phi}_{\widehat{r}_{i'} j}) \right|$, where $\overline{\phi}_{\widehat{r}_i j}$ and $\overline{\phi}_{\widehat{r}_{i'} j}$ are the weighted averages over the q ordinal categories for row cluster $\widehat{r}_i$ and $\widehat{r}_{i'}$ respectively (see eq. (4.1) on page 70).

The $n \times n$ matrix $\{d_{ii'}\}$ describing the crisp clustering can be depicted in a level plot as in Figure 5.13. On the left graph, the students (rows) are shown as they appear in the original data set and, therefore, it is difficult to observe a group pattern. However, the rows were sorted according to the row cluster structure given in (5.1) over both axes on the right contour plot. The crisp cluster structure with $R = 3$ is now easier to identify in this latter graph. A red cell represents that the related two students are allocated to the same cluster, a orange cell depicts a distance of 1 cluster between two students and a yellow cell displays a difference of 2 clusters.

This is a naive representation of the cluster structure but allows us to compare the results between crisp and fuzzy clusterings as we will see below.

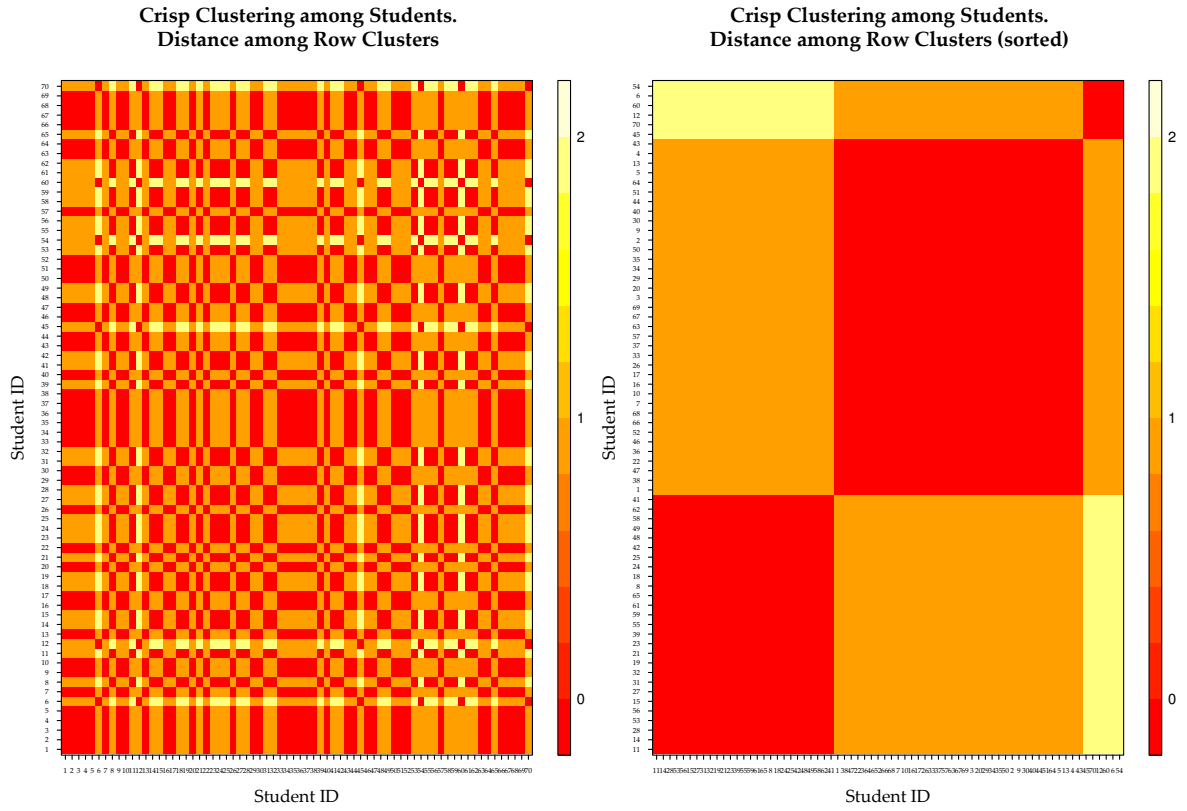

Figure 5.13: ***Applied Statistics course feedback forms data set:*** *Contour plot depicting the crisp cluster structure with $R = 3$ groups. Both axes identify the students (rows). The left figure shows the students without any sorting (i.e. as they appear in the original data set). Both axes are sorted by the row cluster structure given in (5.1) on the right contour plot.*

Fuzzy Clustering Contour Plot

Our approach applies fuzzy clustering via finite mixtures and, therefore, any visualisation tool should take into account any fuzziness in the cluster structure. Figure 5.14 shows two contour plots depicting the probability $C_{ii'}$ of any pair of students i and i' $(i, i' = 1, \ldots, n)$ of being allocated to the same cluster for the Applied Statistics data set. The displayed probability $C_{ii'}$ in both contours is cal-

culated as follows:

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

where $\widehat{Z}_{ir}$ and $\widehat{Z}_{i'r}$ are the posterior probabilities that row i and i' respectively are members of row group r as defined in eq. (2.18) on page 29. It is important to note that we are assuming that the rows are independent conditional on the parameter vector Ω.

On the left graph it is difficult to observe a group pattern because the students are not sorted. However, the right contour plot is sorted by taking into account the row structure given in (5.1) and the $R = 3$ clusters are clearly visible. Red tones represents two students with a high probability of being allocated to the same cluster. Otherwise, orange tones are the students with a moderate probability and yellow tones are those students with lower probability of being allocated to the same cluster. Thus, this pairwise graph of the individuals can depict the cluster structure as the crisp contour plot (Figure 5.13) with the advantage of including the fuzzy assignment of rows to clusters based on the posterior probabilities $\{\widehat{Z}_{ir}\}$.

Score Parameters Distances Level Plot

An alternative way of depicting the fuzziness of the probabilistic clustering is by means of the fitted score parameters. Thus, we can determine the average fitted scores of each row (student) i across all of the m columns (questions) as given in eq. (4.3) in Section 4.2.1:

$$\overline{\phi}_{(i\bullet)} = \frac{1}{m}\sum_{j=1}^{m}\overline{\phi}_{(ij)}, \qquad i = 1, \ldots, n, \tag{5.2}$$

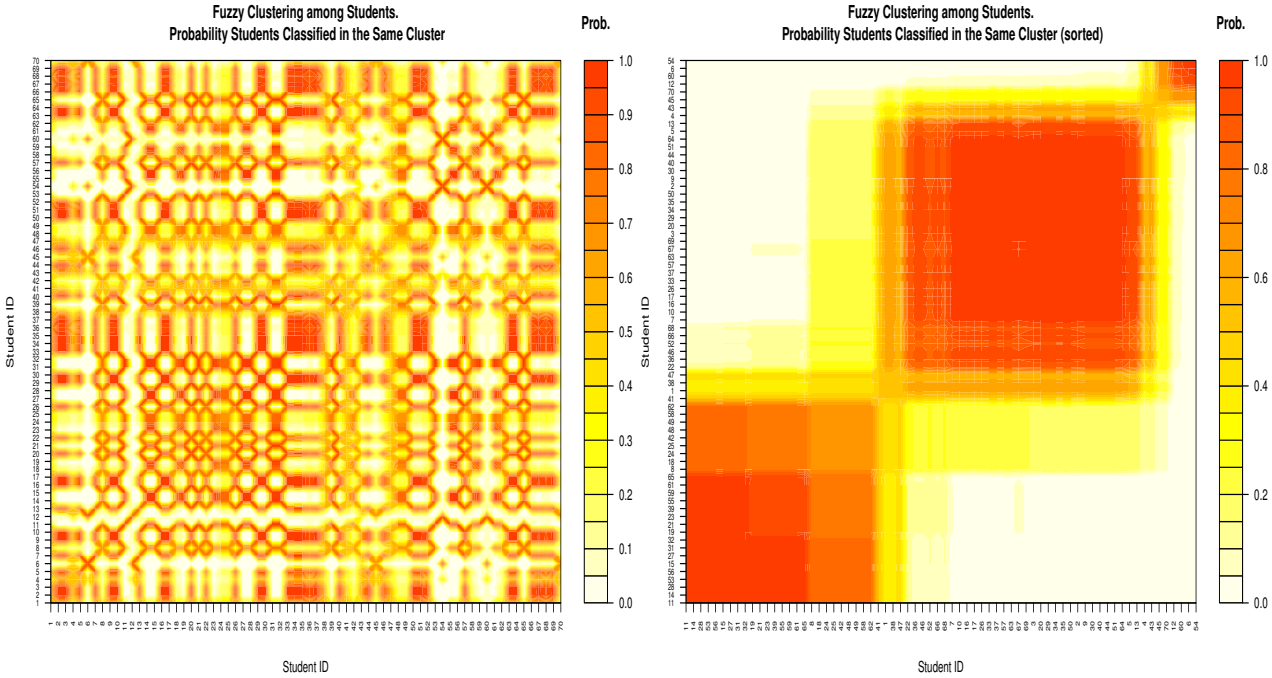

Figure 5.14: ***Applied Statistics course feedback forms data set:*** *Contour plot depicting the fuzzy cluster structure with $R = 3$ groups. Both axes identify the students (rows). The left figure shows the students without any sorting (i.e. as they appear in the original data set). Both axes are sorted by the row cluster structure given in (5.1) on the right contour plot.*

where $\{\overline{\phi}_{(ij)}\}$ is a matrix of the mean response level of each student to each question, conditional on its fuzzy allocation to the $R = 3$ row clusters used above (Section 5.3.3). From here, we can compute the Euclidean distance based on the $\{\overline{\phi}_{(i\bullet)}\}$ values for any two pair of rows (students) so that the differences between the fitted spacing of the levels of the ordinal response can be depicted.

Figure 5.15 presents a side-by-side graph displaying these distances between students. The fuzziness in the clustering is shown using a finest variation cell colour which goes from dark green to light brown. A dark green cell represents two students with a small distance in their fitted scores and therefore very likely to be in the same cluster. A light brown cell depicts high spacing distance between two students and a low possibility of being in the same cluster. The rows were sorted according to the row cluster structure given in (5.1) over both axes on the right graph. As we noted on the fuzzy clustering contour plots (Figure 5.14), the three clusters are easily identifiable on the right level plot.

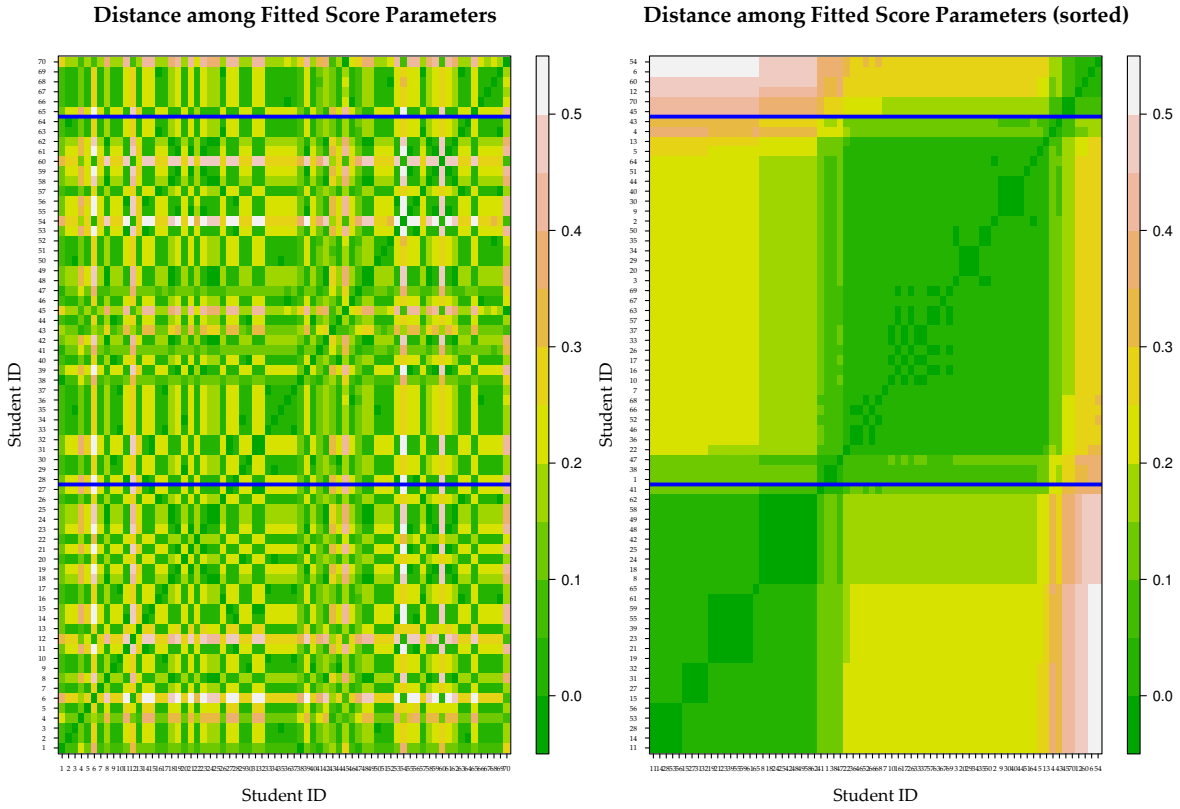

Figure 5.15: ***Applied Statistics course feedback forms data set:*** *Level plots depicting the Euclidean distance in terms of the numerical measure* $\{\overline{\phi}_{(i.)}\}$ *(eq. (5.2)) between a pair of students. Both axes identify the students (rows). The left figure shows the students without any sorting (i.e. as they appear in the original data set). Both axes are sorted by the row cluster structure given in (5.1) on the right contour plot.*

5.4 Discussion

The set of data graphical tools presented in this chapter allows us to easily visualise the results of fitting likelihood-based multivariate methods for data reduction to a matrix containing ordinal data. In particular, these graphs depict the fuzzy probabilistic clustering due to the use of finite mixtures. They are also based on the fitted spacing among levels of the ordinal response variable. This spacing is dictated by the data and arises naturally due to the use of the score parameters $\{\phi_k\}$ from the ordinal stereotype model which are the mixture components from our fuzzy clustering approach.

The visualisation tools presented here all all for the one-dimensional clustering case and have been illustrated with the row clustering version. The graphs

for the column clustering version are essentially the same, but replacing parameters related to rows with the equivalent column parameters. For the case of biclustering, the development of visualisation techniques is a future research direction to explore. One possible direction would be to develop mixture-based biplots as described in Pledger and Arnold (2014). Similar to correspondence analysis, the biplot represents associations among rows, row groups, columns and column groups. Additionally, the spaced mosaic plot introduced in this chapter can be constructed in the case of biclustering. For instance, Figure 5.16 shows a spaced mosaic plot with $R = 2$ row clusters (y-axis) and $C = 2$ column clusters (z-axis) for the spider data set (Section 4.2.3). The description of the graph is the same as explained in Section 5.2.3. The only difference is that we use different colours to differentiate the column boxes within each row box. In this case, blue boxes correspond to column cluster $C = 1$ and the orange ones to column cluster $C = 2$.

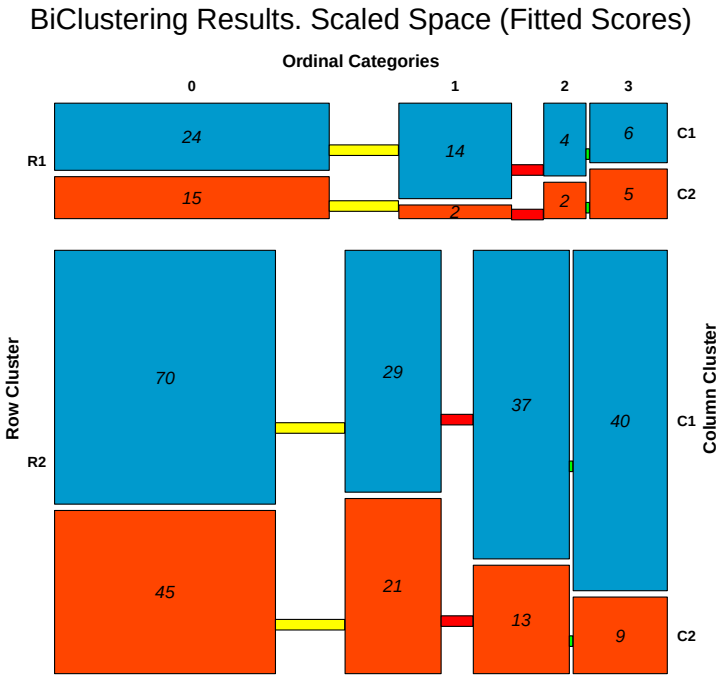

Figure 5.16: ***Spider data set:*** *Mosaic plot with spacing for the biclustering model with $R = 2$ row clusters and $C = 2$ column clusters.*

Chapter 6

Categorising Count Data

6.1 Count data. Description

One of the most common types of data recorded is a count of the number of times an event occurs. A count variable is a type of variable in which the observations arise from counting rather than ordering and take only values on the set of non-negative integers $\{0, \mathbb{Z}^+\}$. The zero value is included in the set of possible values because it possesses a unique and non-arbitrary meaning. A count is a frequency, the number of occurrences of a particular event. The counts may have no upper bound, or may have a known maximum (as in a binomial or multinomial distribution of n objects over different categories). In this chapter, unbounded count data are considered (e.g. observed number of a species in a given area).

Rogers (1974, Chapter 1) describes a stochastic scheme for classifying count data in relation to its variance-mean ratio. When this ratio is equal to unity, i.e. the variance is equal to the mean, the dispersion of the data relative to a predefined study region follows a *random* point pattern (a Poisson process). On the other hand, if the data have a variance-mean ratio greater than unity, i.e. variance>mean (overdispersion), this indicates a more *clustered* (e.g. spatial or temporal clustering) than random point pattern. Finally, if the data has a variance-mean ratio less than unity, i.e. variance<mean (underdispersion), the point pattern is more likely to result from a more *regular* than random or clustered process. In the case of count data distributed as a *random* point pattern, the dispersion is expected to follow a Poisson distribution as the variance of this distribution is equal to its mean. Rogers (1974, Chapter 2) derives the densities under linearity

assumptions detailed below when the dispersion follows a clustered or regular pattern. This is determined in mathematical terms from a random point pattern resulting in a negative binomial distribution when the dispersion is clustered and in a binomial when the dispersion follows a regular pattern. Figure 6.1 illustrates

a. Clustered b. Random c. Regular

Figure 6.1: ***Count data****: Clustered (negative binomial), random (Poisson) and regular (binomial) spatial point patterns over the same region (from Lee and Wong (2001)). Note that the clustered pattern in this particular case is depicted with only one centre but it might also be depicted with more than one centre.*

these point patterns in 3 graphs. The left graph depicts a clustered point pattern (variance>mean, negative binomial distribution) where the probability of an object settling in a quadrat is positively linearly related to the number of objects already there (e.g. shoal of sardines). If this probability is completely independent of the number of objects already in a quadrat (e.g. plants with well-dispersed seeds) then the point pattern is random (variance=mean, Poisson distribution) as shown in the middle graph. The right graph shows a regular point pattern (variance<mean, binomial distribution) where the probability of an object settling in a quadrat decreases linearly with the number of objects already there (e.g. gannet nests in a colony). The mean-variance relationship is a critical property of count data. When not properly controlled for, trends in location (mean abundance) may be confounded with changes in dispersion (variance), leading to misleading results (Warton et al., 2012). One way to deal with the variance-mean ratio problem is to turn the count data into ordinal data. This and other advantages of the use of ordinal data are listed in the next section.

6.2 Advantages of Using Ordinal Data

There are several advantages of categorising an original count data set into ordinal categories. Firstly, one of the causes of overdispersion in count data (variance $>$ mean, see Section 6.1) is the presence of outliers. An estimation approach for count data based on Poisson distributions may be highly sensitive to outliers and produce biased estimates, i.e. the standard errors of the estimates might be deflated or underestimated (Hilbe, 2008, Chapter 4). An ordinal variable is less sensitive to the presence of outliers and therefore the ordinal stereotype model (among other models such as the negative binomial model (NB) and unconditional fixed effects) is better for handling overdispersion.

Secondly, count data is often used on data sets that structurally exclude zero counts (e.g. hospital length of stay data set or number of items in a customer's basket at a supermarket checkout line). The standard Poisson and NB distributions both include the value zero and therefore they should not be used. There are alternative count models to fit this data such as the zero-truncated Poisson (ZTP) and zero-truncated NB (ZTNB) models. These models adjust the probability functions of the Poisson and NB models so that zero counts are excluded, but the sum of probabilities is still one. Another alternative is fitting an ordinal model to a categorised version of the data. The advantage is that an ordinal scale is not affected by the omission of zeros in the data.

A more frequent situation is count data having an excess of zero counts which are far more that the expected zero counts under NB or Poisson distributional assumptions (e.g. number of captured species of spatially rare or hard to detect species). There are several methods for modeling zero-inflated count data such as the hurdle models (also known as zero altered models) (Mullahy, 1986; Heilbron, 1989) and random effect models of various types. The latter includes a zero-inflated Poisson model (ZIP), which for each observation uses a mixture of a Poisson loglinear model and a degenerate distribution at 0, and a zero-inflated NB model (ZINB), which uses an equivalent mixture probability but with a NB loglinear model. This latter model allows overdispersion relative to the zero-inflated data. However, those models can encounter fitting difficulties if there is zero deflation at any settings of the explanatory variables (Agresti, 2010). An alternative to those methods is to introduce finite mixture models with either the Poisson or the NB model as their components. In that case, the ZIP and ZINB

models might be not needed, as the mixtures may separate off zero counts. Another alternative to those methods is to apply a cumulative link random effects model (Saei et al., 1996) to a transformed count data. Thus, the first category is the zero-inflated outcome and each other count class is a separate outcome, turning the count data into ordinal data. If the response variable can take a large number of count outcomes, then the outcomes are grouped into a set of ordinal ordered outcomes. Agresti (2010) recommends grouping by at least four ordinal categories to avoid a substantial efficiency loss. The advantage of this alternative ordinal approach is to require a single set of parameters for describing effects leading to more parsimonious models than the zero-inflated Poisson and zero-inflated NB models which require separate parameters for the effects.

As we described in Section 6.1, the binomial distribution is a useful model to use when count data has underdispersion (variance<mean). The difficulty in this scenario is the estimation of the number of trials parameter. Rogers (1974, Chapter 4) advocates using the largest frequency observed in the data as a estimate for this parameter but it might be biased due to unobserved data. An alternative is to recode the original data into ordinal categories. The optimal categorisation might be determined as an equal number of observations per category, which is a common practice for the χ^2 goodness-of-fit test. However, there has been research showing the power of a χ^2 test may vary substantially with the number of categories (see e.g. Koehler and Gan (1990)). A comprehensive study to find methods of selecting the optimal number of categories should be undertaken in future research.

Conversion of count data to binary outcomes (e.g. presence/absence data) may be seen as an extreme example of conversion to ordinal data with 2 categories (e.g. $y'_{ij} = 0$ or $y'_{ij} > 0$). Our method retains more information than the conversion to binary, retaining more major features of the data but dealing with the variance-mean ratio problem.

Finally, although different models for count data such as Poisson, NB, ZINB and ZTP may be fitted depending on the data features, a good alternative is to recode the data into ordinal scale to fit our ordinal model approach. It enables the inclusion of all of the different cases in one methodology.

6.3 How Many Ordinal Categories?

One of the questions arising from recoding count data into an ordinal scale is related to determining how many ordinal categories into which the data should be optimally categorised. Agresti (2010, Section 2.5) referred to this issue and gives some guidelines for ordinal category choice based on his experience. Although there is no generally accepted methodology for doing this, multiple strategies in different fields have been developed which deal with the categorisation of continuous data. For example, Kotsiantis and Kanellopoulos (2006) reviewed discretization techniques in machine learning context, Hammond and Bickel (2013) summarised recoding techniques in decision analysis, and Dalenius and Hodges (1959) and Lavallee and Hidiriglou (1988) developed algorithms to create optimal stratum boundaries in sample surveys. These methods may be used when recoding count data into ordinal scale. However, to the best of our knowledge, there has not been a lot of research in this subject. In order theory, this recoding is an example of ranking a partially ordered set (count data) into a non-strict weak order or total preorder relation (ordinal scale), where groups of items are formed, and the groups are ordered (see e.g. Ehrgott (2006, Section 1.4) and Roberts and Tesman (2011, Section 4.2.4)).

There are several ways to recode count data into ordinal responses. The simplest case is using the count data as ordinal categories, without collapsing any set of values. For example, recoding a count data set with values $(0, 1, 2, 3)$ as an ordinal variable with label values $\{0, 1, 2, 3\}$. This case would only be tractable if we have data which are concentrated on a small range of counts, and that every count in that range is represented. This direct categorisation implies that although the ordering property is preserved, the property of equal spacing that the counts actually have is removed. A simple extension of this recoding is to categorise large counts into a top-coded data set framework, i.e. data values above an upper bound are censored. This coding is very common in economic surveys (see, for example, the development of regression models to deal with top-coded data sets in Tobin (1958)). For example, we can delimit the values above 2, $(0, 1, 2+)$, and then treat the data values in an ordinal scale $\{0, 1, 2+\}$. An extreme case of top-coding is to dichotomize the count data into just the ordinal scale $\{0, 1+\}$ which converts the data into binary data such as presence-absence data. This implies information loss because the ordinal scale variables tell us more about interspecific

relationships than simple binary data. An alternative is to have equally spaced cut points in counts (e.g. 0-4, 5-9, etc), or equally spaced in counts but on the logarithmic scale (e.g. 0, 1-9, 10-99, 100-999, etc). This practice is common in the formation of strata in sample surveys, where a variable is used to cut a population into mutually exclusive, ordered subgroups. Another option is replacing the count data by their ranks, and then cutting the ranks into groups based on percentiles. This creates an ordinal scale variable.

We have used this latter approach because percentiles are not strongly influenced by extreme values in the count data, and can be calculated even if the counts are skewed. Therefore, percentiles do not depend on the variance-mean ratio scheme of the count data. When recoding a matrix $\boldsymbol{Y} = \{y_{ij}\}$, one option is to recode across the whole count data set with the chosen criterion. However, it may be more appropriate to analyse count data sets where the columns (or rows) have a dramatically different count pattern. For instance in an ecological community, a data set of abundance of species (columns) by sites (rows) might have a set of species with a high count pattern because they are easily detectable species, whereas the rest of species (more difficult to observe) have a low count pattern. This might occur if two species are in competition for the same resources in a particular site, so presence (abundance) of one lowers probability (abundance) of the other. An example of this pattern is observed in the spider data set (Section 4.2.3). In order to obtain a similar frequency distribution for each species, a recoding strategy where the columns are recoded separately. Additionally, as we explain in Section 2.1.2, an advantage of using the ordinal stereotype model in our mixture approach is the interpretation of the score parameters $\{\phi_k\}$. If two ordinal categories have the same (or very similar) score parameter values, this provides evidence that those ordinal categories are not distinguishable and we can collapse them into a single category in our data. It is useful to know into how many cuts (i.e. into how many ordinal categories) the data must be divided.

Given a $n \times m$ matrix $\boldsymbol{Y}$ of count data, our strategy to categorise $\boldsymbol{Y}$ is as follows:

1. Start by setting a large number of ordinal categories q (e.g. $q = 10$).

2. Rescale each observation y_{ij} $(i = 1, \ldots, n, j = 1, \ldots, m)$ as:

$$y_{ij}^{\text{st}} = \frac{y_{ij} - \min(\boldsymbol{Y}_j)}{\max(\boldsymbol{Y}_j) - \min(\boldsymbol{Y}_j)},$$

where $\boldsymbol{Y}_j$ ($j = 1, \ldots, m$) is the column vector.

After this step, we have a new standardized $n \times m$ data matrix $\boldsymbol{Y}^{\text{st}} = \{y_{ij}^{\text{st}}\}$ which lies on the range $[0, 1]$.

3. Divide each new column vector $\boldsymbol{Y}_j^{\text{st}}$ into $q + 1$ quantiles: $Q^{(0)}, \ldots, Q^{(q)}$. There is a number of equivalent ways of defining the sample quantiles. However, the sample quantiles used in statistical packages in common use such as **R** are all based in one or two order statistics, and can be written as:

$$Q^{(k)} = \begin{cases} 0 & \text{if } k = 0, \\ (1 - \varphi) y_{(i)j}^{\text{st}} + \varphi y_{(i+1)j}^{\text{st}} & \text{if } k = 1, \ldots, q - 1, \\ 1 & \text{if } k = q, \end{cases} \tag{6.1}$$

where $\varphi = nk + s - j$, $s = \frac{1}{3}(k + 1)$, $j = \lfloor kn + s \rfloor$ is the floor function for $kn + s$ (i.e. the largest integer not greater than $kn + s$), and $y_{(i)j}^{\text{st}}$ denotes the ith order statistics of the column vector $\boldsymbol{Y}_j^{\text{st}}$ (see Hyndman and Fan (1996) for more details).

4. Recode each observation y_{ij}^{st} ($i = 1, \ldots, n$, $j = 1, \ldots, m$) as:

$$y'_{ij} = \begin{cases} 0 & \text{if } y_{ij}^{\text{st}} = 0, \\ k & \text{if } y_{ij}^{\text{st}} > 0 \text{ and } y_{ij}^{\text{st}} \in (Q^{(k-1)}, Q^{(k)}], \end{cases} \tag{6.2}$$

where $(Q^{(k-1)}, Q^{(k)}]$ is the interval of values from vector $\boldsymbol{Y}_j^{\text{st}}$ between the $(k-1)^{\text{th}}$ and k^{th} quantiles, for $k = 1, \ldots, q$. Each interval contains $\frac{100}{q}\%$ of the non-zero data.

As a result of this step, we obtain an ordinal view $\boldsymbol{Y'}$ of the original data set $\boldsymbol{Y}$.

A graphical illustration of the recoding from count data y_{ij} into ordinal responses y'_{ij} based on the quantiles is given in Figure 6.2.

5. Fit our ordinal mixture methodology to $\boldsymbol{Y'}$.

6. If two or more adjacent categories have the same score parameter value,

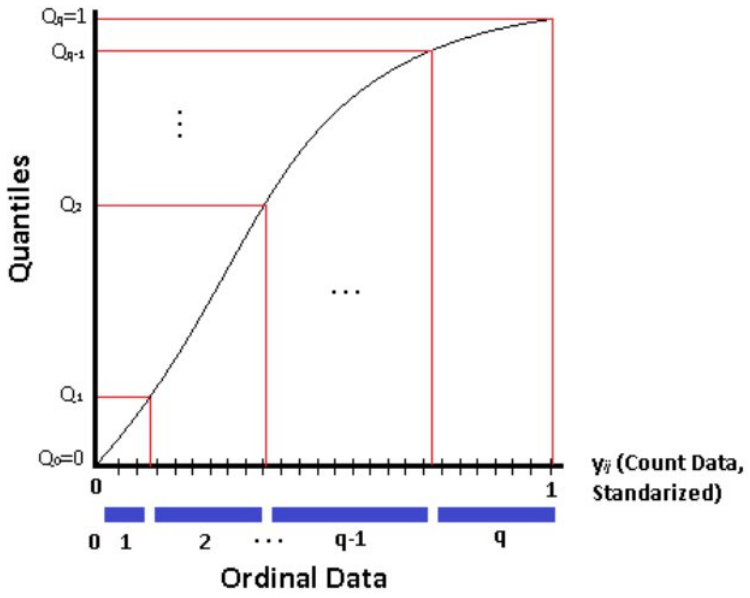

Figure 6.2: ***Quantiles**: Plot illustrating the recoding from count data y_{ij} into ordinal responses y'_{ij} (see eq. (6.2)) based on the quantiles. The y-axis are the quantiles computed as eq. (6.1). The x-axis are the original count data y_{ij} (after standardizing). The red lines divide the plot over the x-axis and y-axis. Each interval between red lines in the y-axis is $(Q^{(k-1)}, Q^{(k)}]$ and creates an interval in the count data (x-axis). The violet blocks are the recoded observation y'_{ij} (eq. (6.2)).*

collapse them, set the new number of ordinal categories q and return to step 2. Otherwise, the categorisation is appropriate and returns the results of model fitting.

Note that we standardize the original count data with the aim of reducing the number of quantiles to calculate in the step 3. Thus, we need to calculate only $(q+1)$ quantiles for the whole data set $\boldsymbol{Y}^{\text{st}}$, instead of $m \times (q+1)$ quantiles (i.e. $q+1$ quantiles for each column in $\boldsymbol{Y}$). However, this standardization might not work suitably for some data sets (e.g. when there is no variation in a column and so the maximum and minimum values in that column are the same) and other strategies can be used. For instance, computing the $q+1$ quantiles for groups of columns. Additionally, we may wish to directly assign zero values from the original count data into a particular category in the ordinal scale (see eq. (6.1)). The reason for this procedure is related to the particular meaning of the zero value in some data sets such as ecological community data regarding species abundance, where it is important to keep absences separated from presences. However, this category

could be removed and equation (6.1) would simply turn into

$$y'_{ij} = \begin{cases} 0 & \text{if } y_{ij}^{\text{st}} \in [Q^{(0)}, Q^{(1)}], \\ \\ k & \text{if } y_{ij}^{\text{st}} \in (Q^{(k)}, Q^{(k+1)}], \end{cases}$$

for $k = 1, \ldots, q-1$. Finally, this strategy was presented on categorising throughout columns but the same idea might be applied over the rows just exchanging columns for rows above.

6.4 Comparing Clusterings

6.4.1 Definition of Measures

The aim of our work is to compare clustering results between count and categorised ordinal data. To the best of our knowledge, no research has been conducted in this field. However, multiple measures have been developed which deal with comparing clusterings over the same data set (see e.g. Strehl and Ghosh (2002), Meila (2005, 2007) and a review in Vinh et al. (2010)). The recoded data set as a result of categorising the original count data into ordinal might be considered different from the original one because some information is not retained in the ordinal scale. Nevertheless, the first stage from our ordinal clustering approach is to take the original data set, generate an ordinal scale and obtain a clustering later on. Therefore, this approach just changes the way in which the data is summarised, and is in some sense similar to analysing continuous data by non-parametric methods using ranks. Additionally, there have been only a few attempts in the literature to develop measures for comparing clusterings over different data sets. They are based on pattern comparison methodologies in data mining (see e.g. Ntoutsi et al. (2006) and Bartolini et al. (2009)) which are beyond the scope of this thesis. Therefore, measures for comparing clusterings in the same data set are considered in this chapter.

Let Y be a data set of N observations, then $\boldsymbol{U} = \{U_1, \ldots, U_K\}$ is a partition, or clustering, of Y into K groups, where $\bigcup_{k=1}^{K} U_k = Y$ and $U_i \cap U_j = \varnothing$ for $i \neq j$ (non-overlapping). Equivalently, $\boldsymbol{V} = \{V_1, \ldots, V_{K'}\}$ on Y is an alternative of clustering Y into K' groups. The information on the overlap between these two clusterings

$\boldsymbol{U}$ and $\boldsymbol{V}$ can be summarised in form of a $K \times K'$ contingency table as illustrated in Table 6.1. Given two clusterings **U** and **V**, the following quantities are defined via the marginal and the joint distributions of data items in **U** and **V** respectively as (Vinh et al., 2010):

$$\begin{aligned}
\mathrm{H}(\mathbf{U}) &= -\sum_{i=1}^{K} \frac{a_i}{N} \log\left(\frac{a_i}{N}\right), && \text{(Entropy for } \mathbf{U}\text{)} \\
\mathrm{H}(\mathbf{V}) &= -\sum_{j=1}^{K'} \frac{b_j}{N} \log\left(\frac{b_j}{N}\right), && \text{(Entropy for } \mathbf{V}\text{)} \\
\mathrm{H}(\mathbf{U},\mathbf{V}) &= -\sum_{i=1}^{K}\sum_{j=1}^{K'} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}}{N}\right), && \text{(Joint entropy for } \mathbf{U} \text{ and } \mathbf{V}\text{)} \\
\mathrm{I}(\mathbf{U},\mathbf{V}) &= \sum_{i=1}^{K}\sum_{j=1}^{K'} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}/N}{a_i b_j/N^2}\right) \\
&= \mathrm{H}(\mathbf{U}) + \mathrm{H}(\mathbf{V}) - \mathrm{H}(\mathbf{U},\mathbf{V}), && \text{(Mutual information for } \mathbf{U} \text{ and } \mathbf{V}\text{)}
\end{aligned} \tag{6.3}$$

where n_{ij} is interpreted as the number of observations from Y that are common to clusters U_i and V_j (i.e. $n_{ij} = |U_i \cap V_j|$), a_i is the sum of row i (i.e. $a_i = |U_i|$), and b_j is the sum of column j (i.e. $b_j = |V_j|$).

Table 6.1: The contingency table for clusterings **U** and **V** on Y where n_{ij} is interpreted as the number of observations from Y that are common to clusters U_i and V_j (i.e. $n_{ij} = |U_i \cap V_j|$), a_i is the sum of row i (i.e. $a_i = |U_i|$), and b_j is the sum of column j (i.e. $b_j = |V_j|$).

$\mathbf{U} \setminus \mathbf{V}$	V_1	V_2	$\ldots$	$V_{K'}$	Total
U_1	n_{11}	n_{12}	$\ldots$	$n_{1K'}$	a_1
U_2	n_{21}	n_{22}	$\ldots$	$n_{2K'}$	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_K	n_{K1}	n_{K2}	$\ldots$	$n_{KK'}$	a_K
Total	b_1	b_2	$\ldots$	$b_{K'}$	$N = \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}$

We use three measures in common use to compare clusterings: the adjusted Rand Index (ARI, Hubert and Arabie (1985)), the variation of information (VI, Meila (2005)), and the normalized information distance (NID, Kraskov et al. (2005)). In the same manner that it is not possible to define a "best" clustering method out of context, one cannot define a measure for comparing clusterings that fits every

problem optimally (Meila, 2007). The ARI is a pair counting-based measure developed from the Rand index (Rand, 1971) and corrected for chance as suggested by Hubert and Arabie (1985). The ARI remains the most well-known and widely used measure to compare clusterings. For instance, Žiberna et al. (2004) used this measure to compare clusterings for ordinal data. The formulation of the ARI from Table 6.1 is as follows,

$$\mathrm{ARI}(\mathbf{U},\mathbf{V}) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K'}\binom{n_{ij}}{2} - \frac{\left[\sum_{i=1}^{K}\binom{a_i}{2}\right]\left[\sum_{j=1}^{K'}\binom{b_j}{2}\right]}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_{i=1}^{K}\binom{a_i}{2} + \sum_{k=1}^{K'}\binom{b_j}{2}\right] - \frac{\left[\sum_{i=1}^{K}\binom{a_i}{2}\right]\left[\sum_{j=1}^{K'}\binom{b_j}{2}\right]}{\binom{N}{2}}}.$$

This measure is bounded above by 1. A 0 value indicates independent clusterings and a 1 value indicates perfect agreement between clusterings.

An alternative to pair counting-based measures (such as ARI) are information theoretic-based distance measures. They are based on the relationship between an observation from Y and its cluster in each of the two clusterings that are compared. Based on the quantities defined in (6.3), the VI for clustering $\mathbf{U}$ and $\mathbf{V}$ is formulated as

$$\mathrm{VI}(\mathbf{U},\mathbf{V}) = \mathrm{H}(\mathbf{U},\mathbf{V}) - \mathrm{I}(\mathbf{U},\mathbf{V}) = 2\mathrm{H}(\mathbf{U},\mathbf{V}) - \mathrm{H}(\mathbf{U}) - \mathrm{H}(\mathbf{V}).$$

This measure is bounded between 0 and $\log(N)$. In order to bound it between 0 and 1, the normalized VI (NVI, Kraskov et al. (2005)) is defined, which consists of dividing $\mathrm{VI}(\mathbf{U},\mathbf{V})$ by $\mathrm{H}(\mathbf{U},\mathbf{V})$:

$$\mathrm{NVI}(\mathbf{U},\mathbf{V}) = 1 - \frac{\mathrm{I}(\mathbf{U},\mathbf{V})}{\mathrm{H}(\mathbf{U},\mathbf{V})}.$$

Another distance measure is the NID which is bounded between 0 and 1 and formulated as

$$\mathrm{NID}(\mathbf{U},\mathbf{V}) = 1 - \frac{\mathrm{I}(\mathbf{U},\mathbf{V})}{\max\{\mathrm{H}(\mathbf{U}),\mathrm{H}(\mathbf{V})\}}.$$

A zero value indicates that $\boldsymbol{U}$ and $\boldsymbol{V}$ are exactly the same clusterings and a value of one is interpreted as independent clusterings for both NVI and NID. Thus, we use the unit-complements of these measures (i.e. 1-NVI and 1-NID) in our comparisons in order to have the same scale interpretation between ARI, NVI and NID. Vinh et al. (2010) advocates that NID is the best measure in widespread

use after analysing the most common distance measures because NID possesses concurrently several important properties, such as using the nominal $[0, 1]$ range better than other normalized distance measures, and satisfying properties of a true metric (i.e. positive definiteness, symmetry and triangle inequality).

6.4.2 Example

In this section, three clusterings from approaches for count and ordinal data are compared. The count data-based clusterings are obtained by applying the likelihood-based methodology described in Pledger and Arnold (2014) for basic Poisson and NB building blocks, and the ordinal data-based clustering are fitted to our ordinal approach. The spider data set (Van der Aart and Smeenk-Enserink, 1974) which was described in Section 4.2.3 is used to compare the three clusterings. This data set shows the distribution of 12 different spider species across 28 different sites. The original count data is shown in Table C.10 in Appendix C.4. Note the large number of zeros and also the high counts, suggesting the NB model is preferable to the Poisson model (Hui et al., 2014). Additionally, Figure 6.3 depicts the variance-mean ratio for all of the species of spiders throughout the sites. The variance is greater than the mean in all the species indicating possibly overdispersion. We categorised the original data into four ordinal responses (see in eq. (4.5) by following the strategy described above (Section 6.3) and the ordinal data set is shown in Table C.11 in Appendix C.4).

According to AIC, the best ordinal data-based clustering was including 3 site groups (see Section 4.2.3 for details). For all the sites, the highest posterior probability stands out from the other two probabilities except for the sites 16, 17 and 19 (e.g. $\kappa_1 = 0.52$ and $\kappa_3 = 0.42$ for site 17). The clustering which allocates the sites 16, 17 and 19 to their highest a posteriori probability cluster is thus not the only reasonable crisp clustering. For this reason we make an alternative allocation ("Stereotype 2") which allocates site 17 to cluster C1 and sites {16, 19} to cluster C2 (whereas they had all been originally allocated to cluster C3). This enable us to test for the effect of the fuzziness when comparing clusterings. Furthermore, we obtained the count data-based clustering for 3 site groups for Poisson and NB building blocks, using the highest probability-based allocation criteria. All the three clusterings are summarised in Table 6.2 and shown in Figure 6.4.

Taking into account the "Stereotype 2" clustering, the results show that sites

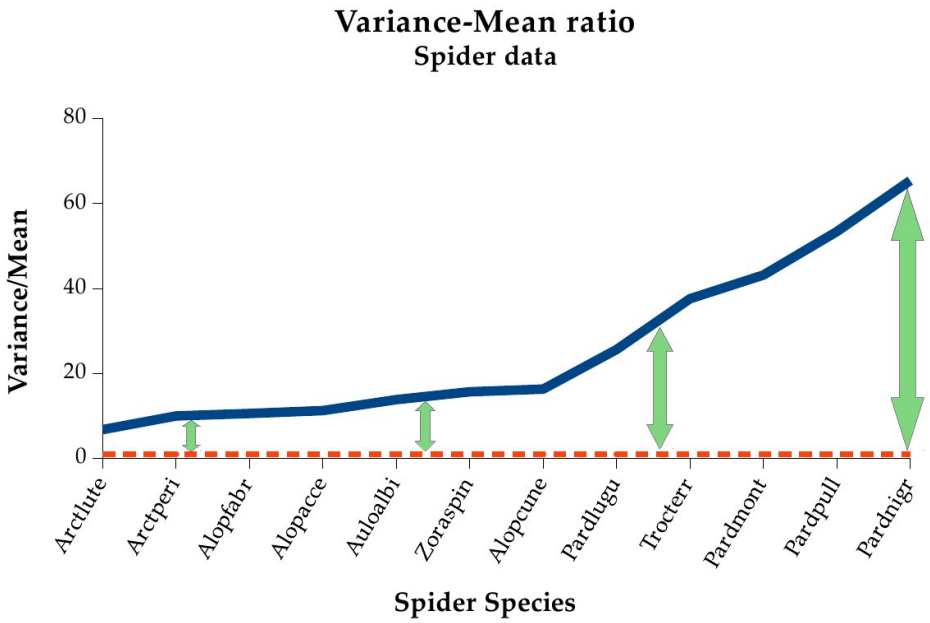

Figure 6.3: ***Spider data set**: Variance-mean ratio (sorted ascending) for the 12 spider species over the sites. The blue lines depicts the amount* $\frac{\text{variance}}{\text{mean}}$ *for each species and the orange dashed line indicates the threshold for no overdispersion* $\left(\frac{\text{variance}}{\text{mean}} = 1\right)$*. Overdispersion (variance>mean) is observed in all the species. The green arrows indicate the magnitude of the overdispersion in each species.*

Table 6.2: **Spider data set**: Clustering results for Poisson, NB and ordered stereotype model. The number of fitted clusters is $C = 3$. All the allocations are based on highest posterior probabilities except for the "Stereotype 2" clustering which has a fuzzy allocation in the sites shown in boldface.

Groups	Clustering (highest probability)			Stereotype 2
	Poisson	NB	Stereotype	
C1	{1-7,9-14,25}	{1-7,13,14}	{1-7,13,14}	{1-7,13,14,**17**}
C2	{22-24,26-28}	{9-12,22-28}	{8,21-24,27,28}	{8,**16**,**19**,21-24,27,28}
C3	{8,15-21}	{8,15-21}	{9-12,15-20,25-26}	{9-12,15,18,20,25-26}

$\{1-7, 13-20, 22-24, 27, 28\}$ are classified into the same cluster for all three probability models. Sites 8 and 21 are allocated to group C2 according to the ordinal model and in group C3 according to the other two models. The opposite happens in site 26. The rest of the sites ($\{9-12, 25\}$) are classified into a different cluster depending on the fitted model.

We want to compare the clustering not only graphically but also using the measures described in Section 6.4. The measures ARI, NVI, and NID were com-

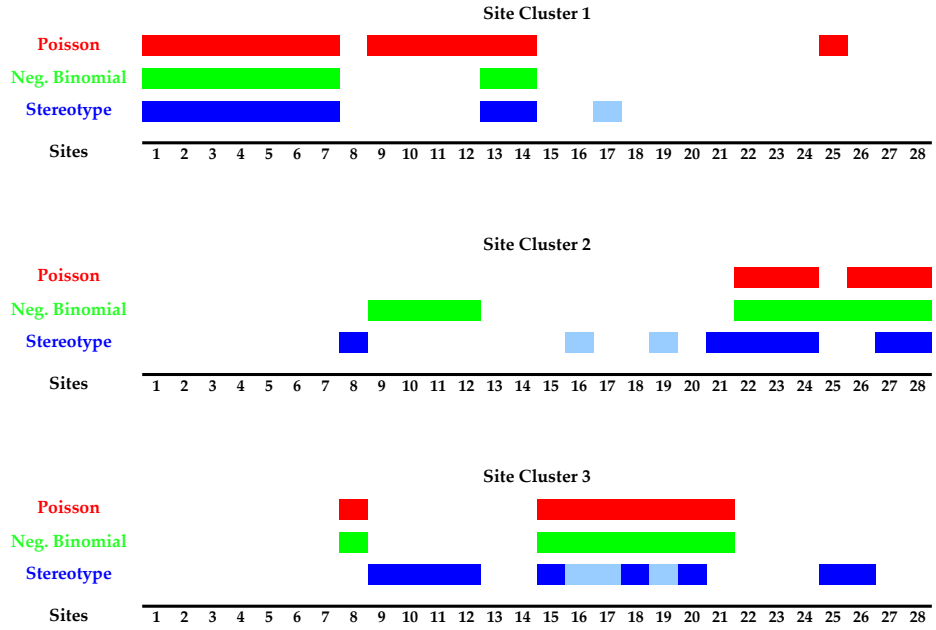

Figure 6.4: ***Spider data set**: Comparison among Poisson (red blocks), NB (green blocks) and ordered stereotype (dark blue blocks) models for the $C = 3$ fitted spider site clustering. The light blue blocks are related to "fuzzy" sites which could classified in more than one cluster.*

puted for the three clusterings (Poisson, NB and Stereotype) and the results are summarised in the Table 6.3. For the three comparison measures, the Poisson

Table 6.3: **Spider data set**: Clustering results for Poisson, NB, and two classifications based on the ordered stereotype model ("Stereotype" and "Stereotype 2"). The number of fitted clusters is $C = 3$. Large values indicate similarity of clustering. The closest clusterings are the two count data-based models (Poisson and NB) over the three measures. Between count and ordinal data-based models, "Stereotype 2" is closer to NB than Poisson and is shown in boldface.

Clustering Comparison	ARI	1-NVI	1-NID
Poisson vs. NB	0.555	0.562	0.701
Poisson vs. Stereotype	0.280	0.229	0.361
NB vs. Stereotype	0.409	0.335	0.500
Poisson vs. Stereotype 2	0.334	0.304	0.457
NB vs. Stereotype 2	**0.465**	**0.423**	**0.590**

and NB clusterings are the closest as it is expected. Between count and ordinal

data-based models, the "Stereotype 2" clustering is closer to the NB clustering than the Poisson one. The clustering from the other ordinal data-based model ("Stereotype") is also closer to NB than Poisson although less similar than the "Stereotype 2". The observed similarity between NB and stereotype clusterings is a satisfactory result because the data is overdispersed suggesting that NB is preferred over Poisson.

6.5 Discussion

We have shown some features of categorising count data into ordinal data in this chapter. In our view, the main advantage is that by using our approach for ordinal data, we do not have to decide among different models for the data. It enables the inclusion of all of the different cases in one methodology. For example, if a count data set involves overdispersion from a set of species and underdispersion from another set, probably the optimal strategy using the original data would be to fit a NB model for the overdispersed set and a binomial model for the underdispersed one. However, we may fit our ordinal stereotype methodology to both of these without treating the data differently. Additionally, many count data sets have extreme variabilities, for example very high counts and very low counts in ecological community data. Replacing these counts with "medium" and "high" ordinal categories makes the actual counts less influential in the model fitting, giving broad categories which enable us to detect major overall patterns.

Although the results shown in this chapter are promising, a more comprehensive study is required in future research. Possible ways to investigate the differences between recoded and original data may be to set up numerical experiments to compare them across a wide range of scenarios or to analyse other data sets with different type of dispersion. Additionally, another future direction might be to study this issue analytically. For instance, developing a measure to quantify the loss of information due to use of the ordinal categorisation instead of the original count data.

Chapter 7

Inference in the Bayesian Paradigm. Fixed Dimension

As we explained in Chapter 2, this thesis proposes a methodology of likelihood-based models for ordinal data including clustering via finite mixtures to define a fuzzy clustering. The ordinal stereotype regression model is used to formulate the ordinal approach. Matechou et al. (2011) developed biclustering models for ordinal data using the proportional odds version of the cumulative logit model and applying a likelihood-based foundation. In Section 2.5, we developed a model fitting procedure to perform a fuzzy clustering assignment for the ordinal stereotype model using the iterative EM algorithm to estimate the parameters. A Bayesian approach to estimate the parameters is introduced in this chapter.

The basics of a Bayesian inference approach are explained in Section 7.1. The framework to implement the Metropolis-Hastings algorithm to our one-dimensional clustering approach is developed and illustrated with a simulation study and two real-life data examples in Section 7.2. Additionally, the label switching problem, which is a common drawback arising from using mixture models, is described in the same section. Finally, conclusions are described in Section 7.3.

7.1 Bayesian Inference

7.1.1 Introduction

The interest in using Bayesian statistics among statisticians from diverse areas has recently increased and so has the area of using finite mixture models for clustering. A good introduction to Bayesian modeling of finite mixtures was given by Marin et al. (2005), Jasra et al. (2005) and Marin and Robert (2007, Chapter 6). Frühwirth-Schnatter (2006) gave a detailed review of Bayesian methods for finite mixtures. With the growth of computing power in the 1990s, the use of Bayesian estimation methods using a Markov chain Monte Carlo (MCMC) procedure as an alternative to the EM algorithm has become increasingly popular (see e.g. McLachlan and Peel (2000) and Lee et al. (2008)). The MCMC algorithm is used in Bayesian inference as a result of the difficulty of constructing the joint and marginal posterior distributions analytically. There are numerous examples of the use of Bayesian methodology with finite mixtures for continuous data. For example, Richardson and Green (1997) and Fraley and Raftery (2007) considered Bayesian methods for the analysis of univariate normal mixtures. Additionally, Stahl and Sallis (2012) showed some examples with Anderson's Iris and Pearson's Crab datasets. However, there is a lack of development of a Bayesian inference approach with mixture models for ordinal data.

A Bayesian inference procedure can deal with some of the difficulties arising in fitting the ordered version of the stereotype model under a likelihood based approach. An advantage of applying a Bayesian approach to our ordinal mixture model is that an appropriate choice for the prior distribution of the score parameters $\{\phi_k\}$ can incorporate their monotone increasing ordinal constraint, $0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1$, so that it is automatically satisfied when generating posterior distributions. Furthermore, we can compute credible intervals of $\{\phi_k\}$ without any additional effort because the Bayesian paradigm is based on simulation of the posterior distribution of the parameters without the need for large sample approximations. Another advantage is that parameter estimation and model selection methodologies do not depend on the regularity conditions required by the LRT and which are violated in the fitting of finite mixtures (see Sections 1.2.1 and 3.2.1). Additionally, there are general advantages in the use of a Bayesian paradigm such as the possibility of incorporating in the estimation

procedure prior knowledge regarding the parameters, the opportunity to update the estimates as new data is available, and the results are more informative as they include the whole joint posterior distribution of the parameters (see a review of advantages in Wagenmakers et al. (2008, Chapter 9)). A Bayesian procedure might be also seen as a method to overcome the problem of occurrence of singularities in the likelihood function (see the approach given by Fraley and Raftery (2007) where the MLE values are replaced by the mode of the posterior). This latter difficulty is one of the main drawbacks when the EM algorithm is used to find the MLE values (see Section 2.5).

However, a Bayesian mixture modeling approach also has drawbacks to overcome such as the choice of suitable prior distributions because the posterior distributions may be heavily influenced by the choice of prior, the high computational cost in MCMC implementations to converge to the results, and the label switching difficulty during MCMC sampling. This latter drawback arises when the label of the mixture components may be arbitrarily permuted on different iterations creating a lack of identifiability. A complete review of the label switching problem for our approach is shown in Section 7.2.5.

In this thesis, the joint posterior distribution of the parameters is obtained by MCMC simulation (see e.g. a comprehensive review by Gilks et al. (1996), Gamerman and Lopes (2006, Chapters 4-6), and Robert and Casella (2010, Chapters 5-8)). The MCMC sampler draws samples from a target distribution (i.e. the posterior distribution) by generating a chain whose stationary distribution is that distribution. In the next section, we introduce some key factors to consider in MCMC in order to assess its reliability.

7.1.2 Considerations for the Use of MCMC

There are three important factors to consider when we use a MCMC sampler: the starting values to initiate the chain, burn-in, and thinning the chain.

Starting values

Assuming that the chain ultimately converges, the choice of starting values does not have influence in making inferences over the stationary distribution because we only use the values obtained in the iterations after the stationary distribution is reached by the MCMC algorithm. At that point, the chain has been running

enough to lose its dependence on the starting values. However, the selection of starting values may affect the performance of the chain in terms of the time it takes to converge to the stationary distribution. Additionally, there is a risk of convergence to local modes and therefore a dispersion of starting points is required for adequate assessment of convergence.

Any values which would be possible to obtain in a sample from the posterior distribution are good starting values. Therefore, the ideal choice of starting values for initiating the MCMC algorithm would be those sampled directly from the posterior distribution, but this is not possible in our case. Thus, we adopt an ad-hoc method for selecting starting values in which a random sample from the prior distribution is used. Section 7.2.2 describes the selection of the prior distributions for all the parameters. A variety of starting values in parallel is necessary to ensure that the sampler has found the mode of the posterior.

Burn-in

A burn-in period in a MCMC sampler means to discard a certain number of the first draws in order to make the chain less dependent on the starting values and only retain the draws closer to the stationary distribution. After the burn-in period, the chain generates values that are presumed samples from the target distribution.

There is nothing in MCMC theory that justifies or even motivates the size of the burn-in period because it relies on the mixing time of the chain to the stationary distribution. A possible procedure would be to do series of short preliminary MCMC runs and to select the iteration number where the chain seems to converge to the stationary distribution as the length of the burn-in period (Geyer, 2011, Chapter 1). In Section 7.1.3, we introduce the measures of convergence diagnostics used throughout this chapter. Another suggestion is to use a visual inspection of the MCMC convergence to determine the burn-in period (Gilks et al., 1996). The most common of those visual tools are briefly described in Section 7.1.3.

Thinning

A MCMC sampler might experience slow convergence (poor mixing) in some particular cases when the dependence between draws is high. In Section 7.1.3,

the autocorrelation plot is described which might be useful to detect high autocorrelations in the chain. In order to reduce sample autocorrelations, a common strategy is to thin the chain by only keeping every d^{th} draw ($d > 0$). Thinning is a practical procedure to increase the efficiency of computer storage or plotting time (see e.g. Link and Eaton (2012) for a discussion of circumstances when thinning might be regarded as a reasonable option).

7.1.3 Convergence Diagnostics for Fixed-Dimensional MCMC

There is no general theory to guarantee the convergence of a MCMC sampler to the target distribution after a given number of runs. However, any implementation of a MCMC algorithm must include a convergence test. The interest relies on how well the chain is mixing over the parameter space in order to obtain reliable parameter estimates. A preliminary procedure to test convergence would be to run a number of parallel chains with overdispersed starting values. If all those chains converge to the same target distribution, then the MCMC simulation has converged to the same stationary distribution. Nevertheless, this procedure should not be taken as a formal diagnosis of convergence since all the chains might be stuck at a local maximum instead of at a global maximum. Unfortunately, if this has happened no convergence test can ever detect this fact.

There are several diagnostic tests that can be used to diagnose convergence. There have been considerable developments in this field in the literature and most of them are based on visual inspections and statistical tests. Common visual inspection tools for the MCMC output are:

- *Trace plot*: It is a plot of the iteration number versus sampled estimations for a parameter at each iteration in the chain. Thus, this graph can be useful to visually check whether the chain has a good mixing in its convergence towards the posterior distribution (i.e. the chain is dense, and remains stable for a long period of time).

- *Probability density function plot*: This plot displays the distribution of the estimations for a parameter during the runtime of the chain. Basically, it is the smoothed histogram of the estimates from the trace plot, i.e. the distribution of the estimates of the parameter in the chain. Convergence is not

easily assessed by this. However, a very noisy highly multimodal plot (i.e. many minor modes) is an indication of convergence failure.

- *Cumulative quantile plot*: It is a plot of the evolution of the sampled quantiles for a estimated parameter as a function of the number of iterations. A sign of a chain with a satisfactory convergence is when each monitored quantile remains stable for a long period of time.

- *Autocorrelation plot*: It shows the autocorrelation function of the chain for an estimated parameter as a function of the number of iterations. High autocorrelations within the chain indicate slow mixing. Moreover, this plot might also be useful as a sign that higher variance proposals would be more efficient when the chain shows high autocorrelations (see Section 7.1.2).

- *Correlation matrix plot*: Since multiple parameters are estimated, we are interested in inspecting the correlation between them. This plot provides an image of the cross-correlation matrix between the estimated parameters. This plot does not inform about convergence, but may assist in constructing a more efficient sampler if a one at a time update strategy is used.

In addition to the visual inspections, convergence should be tested by more formal statistical techniques. Throughout this thesis, we use four of the most common statistical tests in the literature: Geweke time series diagnostic (Geweke, 1992), Gelman and Rubin's multiple sequence diagnostic (Gelman and Rubin, 1992; Brooks, 1998), Heidelberger and Welch diagnostic (Heidelberger and Welch, 1983), and effective sample size (ESS) (Kass et al., 1998). Technical details and an outline of these methods are described in Appendix F. In brief, the Geweke time series diagnostic is based on the comparison of the means of parameter's posterior distributions from two non-overlapping portions of a single chain by using a test for equality of the means. The Gelman and Rubin multiple sequence diagnostic is based on the comparison of a set of chains drawn with overdispersed starting points. The criterion to assess whether the chain converges contrasts the variance within and between chains. The Heidelberger and Welch diagnostic is based on the Cramér-von Mises test statistic to evaluate the null hypothesis that the values drawn from a chain come from a stationary distribution. This diagnostic consists of two tests: a stationary test and a halfwidth test. The stationary test is an iterative procedure based on removing portions of the chain and the

halfwidth test validates the results from the stationary test. The ESS diagnostic uses the autocorrelation function and trace plots as a measure of how well each chain is mixing. Finally, all these diagnostics are implemented in the `coda` (Convergence Diagnosis and Output Analysis) package which is available in **R** (see details in Best et al. (1996) and Plummer et al. (2006, 2012)).

7.1.4 Selecting Models in Bayesian Paradigm

In the data applications of Chapter 4, we mentioned the use of information criteria such as AIC, AICc, BIC and ICL.BIC in order to apply model selection among different alternatives, i.e. select the type of clustering model which best represents the data (row clustering, column clustering and biclustering) and determines the number of different groups in the data. In this section, we describe an information criterion which is commonly used in a Bayesian approach: the deviance information criterion.

Deviance Information Criterion (DIC)

The DIC (Spiegelhalter et al., 2002) is a useful criterion for selecting models under a Bayesian approach. This criterion is defined as a hierarchical modeling generalization of the AIC and BIC and is based on the average of the deviances over the realizations from the posterior distribution, penalized by the effective sample size (ESS) of the chain. The interpretation of DIC is such as the model with a lower DIC is the one with a higher posterior probability and, therefore, it should be selected.

Assume $p(\boldsymbol{Y}|\Omega)$ is a considered model where $\boldsymbol{Y}$ is the data and Ω is the parameter vector. The formulation of the DIC starts with the definition of the Bayesian deviance $D(\Omega)$ as

$$D(\Omega) = -2\log\left(p(\boldsymbol{Y}|\Omega)\right) + 2\log(g(\boldsymbol{Y})), \tag{7.1}$$

where $g(\boldsymbol{Y})$ is a function of the data only. Spiegelhalter et al. (2002) stated that the function $g(\boldsymbol{Y})$ does not need to be specified because this term will be the same when we compare two models and thus it will be cancelled. In addition, if $\widehat{\Omega}$ represents the posterior mean of the parameter vector then the equation (7.1)

becomes

$$D(\widehat{\Omega}) = -2\log\left(p(\boldsymbol{Y}|\widehat{\Omega})\right),$$

which is defined as the Bayesian deviance of the posterior mean. Furthermore, the effective dimension of the model p_D is defined as

$$p_D = \overline{D}(\Omega) - D(\widehat{\Omega}),$$

which gives a measure of the model complexity based on its effective number of parameters. The term $\overline{D}(\Omega)$ is called the posterior mean of Bayesian deviance and is defined as

$$\overline{D}(\Omega) = E_{\Omega|\boldsymbol{Y}}\left[D(\Omega)\right]. \tag{7.2}$$

It can be estimated by the average deviances over the T posterior draws of the parameter vector $\{\Omega_1, \ldots, \Omega_T\}$ as follows

$$\overline{D}(\Omega) = \frac{1}{T}\sum_{i=1}^{T} D(\Omega_i) = \frac{1}{T}\sum_{i=1}^{T}\left\{-2\log\left(p(\boldsymbol{Y}|\Omega_i)\right)\right\}.$$

Finally, the DIC is defined by the expression

$$\text{DIC} = \overline{D}(\Omega) + p_D = 2\overline{D}(\Omega) - D(\widehat{\Omega}).$$

There is controversy about the use of DIC in finite mixture models as a model-choice criterion, as noted by several contributors to the discussion of Spiegelhalter et al. (2002). DeIorio and Robert (2002) described some possible inconsistencies in the definition of DIC for mixture models. Plummer (2008) affirmed that the posterior mean (7.2) is not a suitable estimate for the model parameters since it lies in between multiple modes of the posterior density. Celeux et al. (2006) explored eight different and natural versions of DIC for missing data problems, including mixture models, but finally were unable to recommend any of them. Therefore, implementation of alternatives to DIC in finite mixture models would be a future research directions to consider (Spiegelhalter et al., 2014).

In the following section, we describe the general structure of the Metropolis-Hastings algorithm which is one of the most common Bayesian MCMC samplers.

7.1.5 Description of the Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm (Metropolis et al., 1953) is a MCMC method for obtaining a set of samples from the *target probability distribution* $p(\Omega)$ for which direct sampling is difficult. This algorithm is often used to simulate multivariate distributions. A comprehensive exposition of this algorithm is given in Siddhartha and Greenberg (1995). Basically, the main idea is the simulation of values from a *proposal distribution* $q(\Omega^*|\Omega)$ for the parameter vector Ω, and then apply an acceptance/rejection step to assure the accepted samples come from $p(\Omega)$.

The Metropolis-Hastings algorithm for sampling from a Bayesian posterior distribution $p(\Omega|\boldsymbol{Y})$ is made up of the following iterative steps:

1. Specify an arbitrary initial value for the parameter vector $\Omega^{(0)}$ for which $p(\Omega^{(0)}|\boldsymbol{Y}) > 0$ where $p(\cdot|\boldsymbol{Y})$ is the posterior distribution and $\boldsymbol{Y}$ is the observed data set.

2. Repeat for $t = 1, 2, \ldots, T$

   - Generate a new value for the parameter vector Ω^* from the candidate-generating density (proposal distribution) $q(\Omega^*|\Omega^{(t-1)})$.
   - Compute the acceptance ratio as

$$r = \frac{p(\Omega^*|\boldsymbol{Y})q(\Omega^{(t-1)}|\Omega^*)}{p(\Omega^{(t-1)}|\boldsymbol{Y})q(\Omega^*|\Omega^{(t-1)})}. \tag{7.3}$$

     It is important to note that using Bayes' theorem

$$p(\Omega|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\Omega)\pi(\Omega)}{p(\boldsymbol{Y})} = \frac{L(\Omega|\boldsymbol{Y})\pi(\Omega)}{p(\boldsymbol{Y})},$$

     where $L(\Omega|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\Omega)$ is the likelihood function and $\pi(\Omega)$ the prior density for the parameter vector Ω. The marginal data distribution $p(\boldsymbol{Y})$ cancels in the ratio (7.3) and therefore r can be expressed as a sequence of the following three ratios:

$$r = \frac{L(\Omega^*|\boldsymbol{Y})}{L(\Omega^{(t-1)}|\boldsymbol{Y})} \frac{\pi(\Omega^*)}{\pi(\Omega^{(t-1)})} \frac{q(\Omega^{(t-1)}|\Omega^*)}{q(\Omega^*|\Omega^{(t-1)})}.$$

We can name the left term as the *likelihood ratio*, the term in the middle as the *prior ratio* and the right term as the *proposal ratio*.

- Compute the probability of move $\alpha(\Omega^*|\Omega^{(t-1)}) = \min\{r, 1\}$
- Generate a value u from a $\mathcal{U}(0, 1)$ distribution.
- If $u \leq \alpha(\Omega^*|\Omega^{(t-1)})$, we accept the new value $\Omega^{(t)} = \Omega^*$. Otherwise, we reject it and set $\Omega^{(t)} = \Omega^{(t-1)}$.

3. Return the values $\{\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(T)}\}$.

4. Test whether convergence has been achieved. If not, increase T and return to step 2.

Although the proposal distribution $q(\cdot|\cdot)$ is essentially arbitrary, it has to be chosen carefully in order to avoid long sampling runs and, consequently, minimize computational time. A recommendation is to choose $q(\cdot|\cdot)$ so that a reasonable proportion of candidates are accepted. One possible option is to choose $q(\cdot|\cdot)$ so that the new candidate point is close to the current point, so that is likely to be accepted. However, if it is too close, the chain will mix very slowly.

Gibbs sampling (Geman and Geman, 1984) is a special case of Metropolis-Hastings sampling where the new value is always accepted (i.e. the probability of move is always 1). Gibbs sampling performs a random walk where random variables are simulated sequentially from univariate conditional distributions rather than from the full joint distribution. The Gibbs sampler only considers univariate conditional distributions where all of the random variables but one are assigned fixed values. Compared with the Gibbs sampler, M-H algorithms can be tuned toward a much wider range of possibilities.

The sampled posterior distribution can be summarised by means of its mean, median, standard deviation (SD), time-series standard error, and 95% highest posterior density (HPD) interval. Given a predetermined level such as $\alpha = 0.05$, the $(1-\alpha)\%$ HPD interval for a parameter θ is a credible interval which is constructed from the marginal posterior distribution of the parameter as the smallest possible length interval for which the difference in the probability values of the interval endpoints is $(1-\alpha)\%$. That is, $\text{HPD}(\boldsymbol{Y}) = \{\Omega; p(\Omega|\boldsymbol{Y}) \geq k_\alpha\}$ where $\boldsymbol{Y}$ is the observed data set and k_α is determined by the coverage constraint $p(\Omega \in \text{HPD}(\boldsymbol{Y})) = 1 - \alpha$. We use this definition throughout this thesis.

7.2 Fixed Dimension: Metropolis-Hastings Sampler

In this section, we develop the methodology to estimate the parameters for our one-dimensional clustering finite mixture-density model with incomplete data (the unknown row (column) membership probability). Sections 7.2.1-7.2.3 describe the components for formulating the Bayesian inference for our approach: the likelihood function (Section 7.2.1), the prior distributions imposed upon score, cut point and row cluster parameters and their corresponding full conditional distributions (Section 7.2.2), and the joint posterior distribution (Section 7.2.3). Finally, Section 7.2.4 describes the parameter estimation procedure by using the M-H sampler. This development is focused on the row clustering model version (the procedure for column clustering model is similar).

7.2.1 Likelihood Function

As explained in Section 2.1.1, for ordinal response variables with q categories, the data are represented by a $n \times m$ matrix $\boldsymbol{Y}$ where for instance the n rows might represent the subjects in a particular questionnaire and the m columns be the different questions, and

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

As formulated on eq. (2.7) on page 25 (Section 2.3), the probability of the data response y_{ij} being equal to category k given that individual i is in row group r $(i \in r)$ including row clustering and interaction factor is

$$\begin{aligned} \theta_{rjk} &= P\left[y_{ij} = k | i \in r\right] = \frac{\exp(\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj}))}{\sum_{\ell=1}^{q} \exp(\mu_\ell + \phi_\ell(\alpha_r + \beta_j + \gamma_{rj}))} \\ r &= 1, \ldots, R, \quad j = 1, \ldots, m, \quad k = 1, \ldots, q, \end{aligned} \tag{7.4}$$

including the monotone increasing constraint $0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1$ in the score parameters and $\mu_1 = 0$ for reasons of identifiability. Identifiability also necessitates imposing $\sum_{r=1}^{R} \alpha_r = \sum_{j=1}^{m} \beta_j = 0$ and sum-to-zero constraints on each row and column of the association matrix $\{\gamma_{rj}\}$, i.e. $\sum_{r=1}^{R} \gamma_{rj} = 0$ for $j = 1, \ldots, m$ and $\sum_{j=1}^{m} \gamma_{rj} = 0$ for $r = 1, \ldots, R$. We can formulate the likelihood

function as

$$L(\Omega|\{y_{ij}\}) = \prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r \prod_{j=1}^{m}\prod_{k=1}^{q}(\theta_{rjk})^{I(y_{ij}=k)}\right], \tag{7.5}$$

where Ω is the parameter vector.

7.2.2 Prior Distributions

There are six sets of parameters in Ω:

- $q-1$ *cut point* parameters $\{\mu_k\}$,
- $q-2$ score parameters $\{\phi_k\}$,
- $R-1$ row clustering effects $\{\alpha_r\}$,
- $m-1$ column effects $\{\beta_j\}$,
- $(R-1)(m-1)$ interaction factors $\{\gamma_{rj}\}$, and
- $R-1$ row group membership probability parameters $\{\pi_r\}$.

Proper prior distributions for each set of parameters are described below. They chose to minimize the impact of the priors on the inference by choosing *noninformative* priors for all parameters, but in situations where prior knowledge is available these can easily be replaced with more concentrated distributions

Cut Point Parameters

In the case of the independent cut point parameters, we impose a normal prior distribution:

$$\mu_k \sim \mathcal{N}(\mu_\mu = 0, \sigma_\mu^2) \qquad (k = 2, \ldots, q \quad \text{and} \quad \mu_1 = 0), \tag{7.6}$$

where σ_μ^2 follows a hyperprior

$$\sigma_\mu^2 \sim \text{InverseGamma}\left(\nu_\mu, \delta_\mu\right),$$

where $\nu_\mu = 3$ and $\delta_\mu = 40$.

Score Parameters

We reparametrise the score parameters in terms of their differences in order to handle with their monotone increasing ordinal constraint $0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1$. Thus, we define a new variable ν_k as $\nu_k = \phi_{k+1} - \phi_k$ for $k = 1, \ldots, q-2$. Given this reparametrisation and the constraint in the score parameters $\{\phi_k\}$, the new parameters $\{\nu_k\}$ all lie in $0 \leq \nu_k \leq 1$. In addition, $\sum_{k=1}^{q-1} \nu_k = 1$ and therefore $\nu_{q-1} = 1 - \sum_{k=1}^{q-2} \nu_k$. Finally, it is easy to prove that the set of original parameters $\{\phi_k\}$ satisfy $\phi_k = \sum_{\ell=1}^{k-1} \nu_\ell$.

We assume that a priori these differences $(\nu_1, \ldots, \nu_{q-1})$ follow a Dirichlet joint distribution with parameters $\boldsymbol{\lambda}_\phi = (\lambda_{\phi_1}, \lambda_{\phi_2}, \ldots, \lambda_{\phi_{q-1}})$ (Ahn et al., 2009):

$$\pi(\nu_1, ..., \nu_{q-1}) = \frac{\Gamma(\lambda_{\phi_1} + \cdots + \lambda_{\phi_{q-1}})}{\Gamma(\lambda_{\phi_1})\Gamma(\lambda_{\phi_2}) \cdots \Gamma(\lambda_{\phi_{q-1}})} \prod_{k=1}^{q-1} \nu_k^{\lambda_{\phi_k}-1}. \tag{7.7}$$

Moreover, we can establish a random variable transformation from $\{\nu_k\}$ to $\{\phi_k\}$ as

$$\pi(\phi_2, ..., \phi_{q-1}) = \pi(\nu_1, ..., \nu_{q-2}) \left| \frac{\partial(\nu_1, ..., \nu_{q-2})}{\partial(\phi_2, ..., \phi_{q-1})} \right| \qquad \left(\nu_{q-1} = 1 - \sum_{k=1}^{q-2} \nu_k \right),$$

where the determinant of the Jacobian matrix is:

$$\left| \frac{\partial(\nu_1, ..., \nu_{q-2})}{\partial(\phi_2, ..., \phi_{q-1})} \right| = \begin{vmatrix} \frac{\partial \nu_1}{\partial \phi_2} & \frac{\partial \nu_1}{\partial \phi_3} & \cdots & \frac{\partial \nu_1}{\partial \phi_{q-1}} \\ \frac{\partial \nu_2}{\partial \phi_2} & \frac{\partial \nu_2}{\partial \phi_3} & \cdots & \frac{\partial \nu_2}{\partial \phi_{q-1}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial \nu_{q-2}}{\partial \phi_2} & \frac{\partial \nu_{q-2}}{\partial \phi_3} & \cdots & \frac{\partial \nu_{q-2}}{\partial \phi_{q-1}} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{vmatrix} = 1.$$

In that manner, we can formulate the joint prior distribution for the score parameters $\{\phi_k\}$ with hyperparameters $\boldsymbol{\lambda}_\phi$:

$$\begin{aligned} \pi(\{\phi_k\}) &= \frac{\Gamma(\lambda_{\phi_1} + \cdots + \lambda_{\phi_{q-1}})}{\Gamma(\lambda_{\phi_1})\Gamma(\lambda_{\phi_2}) \cdots \Gamma(\lambda_{\phi_{q-1}})} \prod_{k=1}^{q-1} \nu_k^{\lambda_{\phi_k}-1} \\ &= \frac{\Gamma(\lambda_{\phi_1} + \cdots + \lambda_{\phi_{q-1}})}{\Gamma(\lambda_{\phi_1})\Gamma(\lambda_{\phi_2}) \cdots \Gamma(\lambda_{\phi_{q-1}})} \prod_{k=1}^{q-1} (\phi_{k+1} - \phi_k)^{\lambda_{\phi_k}-1} \qquad (\phi_1 = 0 \text{ and } \phi_q = 1). \end{aligned} \tag{7.8}$$

If $\{\phi_k\}$ are the order statistics of $q-2$ draws from a uniform distribution $U(0,1)$ with $\phi_1 = 0$ and $\phi_q = 1$, then the successive differences of the order statistics $\{\nu_k\}$ follow a Dirichlet($\boldsymbol{\lambda}_\phi$) distribution with $\boldsymbol{\lambda}_\phi = \mathbf{1}$. This prior is equivalent to imposing an equal expected value over the score parameters.

Row Clustering Effect and Column Effect Parameters

Regarding the R row clustering effect parameters $\{\alpha_r\}$ and the m column effect parameters $\{\beta_j\}$, we impose the following one-dimensional degenerate normal prior distributions:

$$\begin{aligned}(\alpha_1,\ldots,\alpha_R) &\sim \text{DegenNormal}(R;\mu_\alpha=0,\sigma_\alpha^2) \quad \left(\sum_{r=1}^{R}\alpha_r=0\right), \text{ and}\\ (\beta_1,\ldots,\beta_m) &\sim \text{DegenNormal}(m;\mu_\beta=0,\sigma_\beta^2) \quad \left(\sum_{j=1}^{m}\beta_j=0\right),\end{aligned} \tag{7.9}$$

where the parameters σ_α and σ_β follow hyperpriors

$$\sigma_\alpha^2 \sim \text{InverseGamma}\,(\nu_\alpha,\delta_\alpha) \;\text{ and }\; \sigma_\beta^2 \sim \text{InverseGamma}\,(\nu_\beta,\delta_\beta)\,,$$

where $\nu_\alpha = \nu_\beta = 3$ and $\delta_\alpha = \delta_\beta = 40$. The reason to set parameters μ_α and μ_β to zero is because the row clustering effect and column effect parameters are constrained to have a zero sum. In that manner, $\{\alpha_r\}$ and $\{\beta_j\}$ are vectors of draws from a normal distribution:

$$\begin{aligned}\alpha_r &\sim \mathcal{N}\left(0,\frac{R-1}{R}\sigma_\alpha^2\right) \qquad r=1,\ldots,R-1, \text{ and}\\ \beta_j &\sim \mathcal{N}\left(0,\frac{m-1}{m}\sigma_\beta^2\right) \qquad j=1,\ldots,m-1,\end{aligned}$$

constrained to have a zero sum $\left(\alpha_R = -\sum_{r=1}^{R-1}\alpha_r \text{ and } \beta_m = -\sum_{j=1}^{m-1}\beta_j\right)$. Any pair of distinct components from $\{\alpha_r\}$ and $\{\beta_j\}$ are negatively correlated with $\text{Cov}(\alpha_a,\alpha_b) = -\sigma_\alpha^2/R$ and $\text{Cov}(\beta_a,\beta_b) = -\sigma_\beta^2/m$ $(a \neq b)$. Proofs and brief details of the one-dimensional degenerate normal distributions are given in Appendix E.1.1.

Interaction Factor Parameters

The prior distribution for the interaction factor parameters $\{\gamma_{rj}\}$ follows a two-dimensional degenerate normal distribution,

$$\begin{aligned} \{\gamma_{rj}\} &\sim \text{DegenNormal}(R, m; \mu_\gamma = 0, \sigma_\gamma^2) \\ r &= 1, \ldots, R, \quad j = 1, \ldots, m, \end{aligned} \tag{7.10}$$

where the variance σ_γ^2 of the prior follows a hyperprior distribution:

$$\sigma_\gamma^2 \sim \text{InverseGamma}\left(\nu_\gamma, \delta_\gamma\right),$$

where $\nu_\gamma = 3$ and $\delta_\gamma = 40$. In addition, the parameter μ_γ is set to zero so that $\{\gamma_{rj}\}$ is a matrix of draws from a normal distribution, satisfying that on each row and column have zero sum. Any individual component of the matrix $\{\gamma_{rj}\}$ is distributed as:

$$\gamma_{rj} \sim \mathcal{N}\left(0, \frac{(R-1)(m-1)}{Rm}\sigma_\gamma^2\right) \qquad r = 1, \ldots, R, \quad j = 1, \ldots, m,$$

constrained to have a zero sum ($\sum_{r=1}^{R} \gamma_{rj} = 0$ for $j = 1, \ldots, m$ and $\sum_{j=1}^{m} \gamma_{rj} = 0$ for $r = 1, \ldots, R$). Any pair of components from $\{\gamma_{rj}\}$ is correlated as

$$\text{Cov}(\gamma_{rj}, \gamma_{r'j'}) = \frac{\sigma_\gamma^2}{Rm}, \text{Cov}(\gamma_{rj}, \gamma_{rj'}) = -\sigma_\gamma^2\frac{(R-1)}{Rm}, \text{ and } \text{Cov}(\gamma_{rj}, \gamma_{r'j}) = -\sigma_\gamma^2\frac{(m-1)}{Rm}$$

where $i \neq i'$, $j \neq j'$. Proofs and brief details of the two-dimensional degenerate normal distribution are given in Appendix E.1.2.

Row Membership Probability Parameters

We directly impose a Dirichlet distribution on the R parameters corresponding to the unknown row membership probabilities $\{\pi_r\}$

$$(\pi_1, \ldots, \pi_R) \sim \text{Dirichlet}(\boldsymbol{\lambda}_\pi) \qquad \left(\sum_{r=1}^{R} \pi_r = 1\right). \tag{7.11}$$

where $\boldsymbol{\lambda}_\pi = (\lambda_{\pi_1}, \ldots, \lambda_{\pi_R})$ are the hyperparameters. We set $\boldsymbol{\lambda}_\pi = \mathbf{1}$. This prior is equivalent to imposing an equal expected value over the row membership pa-

rameters.

7.2.3 Joint Posterior Distribution

Combining the prior distributions (7.6) and (7.8)-(7.11), and the formulation of the likelihood function (7.5), then we can derive the following joint posterior distribution conditional to the data $\boldsymbol{Y}$ and the number of row clusters R as

$$p(\Omega|\boldsymbol{Y}, R) \propto L(\Omega|\boldsymbol{Y}, R) \times \pi(\{\mu_k\})\pi(\{\phi_k\})\pi(\{\alpha_r\})\pi(\{\beta_j\})\pi(\{\gamma_{rj}\})\pi(\{\pi_r\}), \quad (7.12)$$

where the factor $\pi(\cdot)$ are the prior distributions for each parameter. The priors for this model are listed in Table 7.1, with default values of the relevant defining constants given alongside.

Table 7.1: Metropolis-Hastings sampler. Priors and default settings for the hyperparameters defining their distributions.

Parameter	Prior Distribution	Hyperparameters
σ_μ^2	InverseGamma (ν_μ, δ_μ)	$\nu_\mu = 3$ $\delta_\mu = 40$
$\{\mu_k\}$	$\mathcal{N}(0, \sigma_\mu^2)$	
$\{\phi_k\}$	$\{\nu_k\} \sim$ Dirichlet$(\boldsymbol{\lambda}_\phi)$ $\nu_k = \phi_{k+1} - \phi_k$	$\boldsymbol{\lambda}_\phi = \mathbf{1}$
σ_α^2	InverseGamma $(\nu_\alpha, \delta_\alpha)$	$\nu_\alpha = 3$ $\delta_\alpha = 40$
$\{\alpha_r\}$	DegenNormal$(R; 0, \sigma_\alpha^2)$	
σ_β^2	InverseGamma $(\nu_\beta, \delta_\beta)$	$\nu_\beta = 3$ $\delta_\beta = 40$
$\{\beta_j\}$	DegenNormal$(m; 0, \sigma_\beta^2)$	
$\{\gamma_{rj}\}$	DegenNormal$(R, m; 0, \sigma_\gamma^2)$	$\sigma_\gamma^2 = 5$
$\{\pi_r\}$	Dirichlet$(\boldsymbol{\lambda}_\pi)$	$\boldsymbol{\lambda}_\pi = \mathbf{1}$

7.2.4 Posterior Estimation

The parameter vector to estimate Ω is composed of the q-vectors $\boldsymbol{\phi} = \{\phi_k\}$ and $\boldsymbol{\mu} = \{\mu_k\}$, the R-vectors $\boldsymbol{\alpha} = \{\alpha_r\}$ and $\boldsymbol{\pi} = \{\pi_r\}$, the m-vector $\boldsymbol{\beta} = \{\beta_j\}$ and

the $(R \times m)$ matrix $\boldsymbol{\gamma} = \{\gamma_{rj}\}$. We now use the Metropolis-Hastings algorithm (as described in the Section 7.1.5) to construct a sampler for the joint posterior $p(\Omega|\boldsymbol{Y}, R)$ given an observed data matrix $\boldsymbol{Y}$ and assuming R row clusters. This iterative process is summarised in the following way:

1. Specify arbitrary initial values for the parameter vector $\Omega^{(0)}$. All the initial values for $\Omega^{(0)}$ are drawn from the prior distributions (Table 7.1):

   - **Initial value for $\boldsymbol{\mu}$**: All cut point parameters $\{\mu_k\}$ are a priori independent. Thus, we use the marginal distributions and take a random draw from a univariate normal distribution $\mathcal{N}(0, \sigma_\mu^2)$ as an arbitrary initial value for each element $\mu_k^{(0)}$ $(k = 2, \ldots, q)$. The hyperprior for σ_μ^2 and its hyperparameters are stated in Table 7.1.
   - **Initial value for $\boldsymbol{\phi}$**: We draw a random sample from (7.7):

     $$(\nu_1^{(0)}, \ldots, \nu_{q-2}^{(0)}) \sim \text{Dirichlet}(\lambda_{\phi_1}, \ldots, \lambda_{\phi_{q-2}}),$$

     and then we establish the starting values for $\{\phi_k^{(0)}\}$ by setting $\phi_k^{(0)} = \sum_{\ell=1}^{k-1} \nu_\ell^{(0)}$ for $k = 2, \ldots, q-1$, $\phi_1^{(0)} = 0$ and $\phi_q^{(0)} = 1$.
   - **Initial value for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$**: We draw a random sample from the one-dimensional degenerate normal prior distributions (7.9):

     $$(\alpha_1^{(0)}, \ldots, \alpha_R^{(0)}) \sim \text{DegenNormal}(R; 0, \sigma_\alpha^2),$$
     $$(\beta_1^{(0)}, \ldots, \beta_m^{(0)}) \sim \text{DegenNormal}(m; 0, \sigma_\beta^2),$$

     where $\sum_{r=1}^R \alpha_r^{(0)} = 0$ and $\sum_{j=1}^m \beta_j^{(0)} = 0$. In the case of the row cluster effect parameters $\{\alpha_r\}$, the prior on the parameter σ_α^2 is InverseGamma $(\nu_\alpha, \delta_\alpha)$. The same idea states for the column effect parameters $\{\beta_j\}$ (InverseGamma $(\nu_\beta, \delta_\beta)$). The imposed values for hyperparameters $\{\nu_\alpha, \delta_\alpha, \nu_\beta, \delta_\beta\}$ are given in Table 7.1.
   - **Initial value for $\boldsymbol{\gamma}$**: We draw a random sample from the two-dimensional degenerate normal prior distributions (7.10):

     $$\{\gamma_{rj}^{(0)}\} \sim \text{DegenNormal}(R, m; 0, \sigma_\gamma^2) \qquad r = 1, \ldots, R, \quad j = 1, \ldots, m,$$

     where $\sum_{r=1}^R \gamma_{rj}^{(0)} = 0$ for all j, $\sum_{j=1}^m \gamma_{rj}^{(0)} = 0$ for all r, and a fix value

$\sigma_\gamma^2 = 5$ is taken.

- **Initial value for $\boldsymbol{\pi}$**: We specify a random sample from a Dirichlet distribution with parameter vector $\boldsymbol{\lambda}_\pi = \mathbf{1}$ as starting values for $\{{\pi_r}^{(0)}\}$. This hyperparameter choice minimizes the impact of the prior on the inference for the row cluster membership parameters (noninformative prior).

2. At t^{th} iteration ($t = 1, 2, \ldots, T$):

   (a) We define the likelihood ratio at iteration t (LR) from (7.5) as

$$\text{LR} = \frac{\prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r^*\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\theta_{rjk}^*\right)^{I(y_{ij}=k)}\right]}{\prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r^{(t-1)}\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\theta_{rjk}^{(t-1)}\right)^{I(y_{ij}=k)}\right]}, \tag{7.13}$$

   where θ_{rjk}^* is the probability defined as (7.4).
   Additionally, the prior distribution for parameter vector $\boldsymbol{\alpha}$ ($\pi(\boldsymbol{\alpha})$) is a DegenNormal$(R; 0, \sigma_\alpha^2)$ distribution (see Table 7.1) and therefore the pdf is (see eq. (E.1) in Appendix E.1.1):

$$\begin{aligned}\pi(\boldsymbol{\alpha}) = f_{\text{Deg}\mathcal{N}}(\boldsymbol{\alpha}|R,\sigma_\alpha^2) &= (2\pi\sigma_\alpha^2)^{-\left(\frac{R-1}{2}\right)}R^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma_\alpha^2}\sum_{r=1}^{R}\alpha_r^2\right)\delta\left(\sum_{r=1}^{R}\alpha_r\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_\alpha^2}\sum_{r=1}^{R}\alpha_r^2\right)\delta\left(\sum_{r=1}^{R}\alpha_r\right).\end{aligned} \tag{7.14}$$

   Similarly the formulation of the prior distribution for parameter vector $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta}) = f_{\text{Deg}\mathcal{N}}(\boldsymbol{\beta}|m,\sigma_\beta^2) \propto \exp\left(-\frac{1}{2\sigma_\beta^2}\sum_{j=1}^{m}\beta_j^2\right)\delta\left(\sum_{j=1}^{m}\beta_j\right), \tag{7.15}$$

and the pdf of $\boldsymbol{\gamma}$ is (see eq. (E.5) in Appendix E.1.1):

$$\begin{aligned}\pi(\boldsymbol{\gamma}) = f_{\mathrm{Deg}\mathcal{N}}(\boldsymbol{\gamma}|R,m,\sigma_\gamma^2) &= (2\pi\sigma_\gamma^2)^{-\frac{(R-1)(m-1)}{2}} R^{\frac{(m-1)}{2}} m^{\frac{(R-1)}{2}} \\ &\times \exp\left(-\frac{1}{2\sigma_\gamma^2}\sum_{r=1}^{R}\sum_{j=1}^{m}\gamma_{rj}^2\right)\prod_{r=1}^{R}\delta\left(\sum_{j=1}^{m}\gamma_{rj}\right)\prod_{j=1}^{m-1}\delta\left(\sum_{r=1}^{R}\gamma_{rj}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_\gamma^2}\sum_{r=1}^{R}\sum_{j=1}^{m}\gamma_{rj}^2\right)\prod_{r=1}^{R}\delta\left(\sum_{j=1}^{m}\gamma_{rj}\right)\prod_{j=1}^{m-1}\delta\left(\sum_{r=1}^{R}\gamma_{rj}\right).\end{aligned} \tag{7.16}$$

(b) Generate a new candidate for the parameter Ω^* and test whether we accept it as a new value for the parameter. At each iteration, we update blocks of parameters in turn (e.g. first $\boldsymbol{\phi}^*$, then $\boldsymbol{\mu}^*$, and so on) but within each block we just select a random component, or a neighbouring pair of parameters, to update in order to avoid slow mixing.
The parameter vector $\boldsymbol{\Omega}$ is updated as follows:

- **New candidate for $\boldsymbol{\mu}$**: We update a randomly selected element ℓ of $\boldsymbol{\mu}$ ($\ell \in 2,\ldots,q$). Therefore, the probability that a particular μ_ℓ is selected is $1/(q-1)$. The proposal distribution $q(\boldsymbol{\mu}^*|\boldsymbol{\mu}^{(t-1)})$, is imposed via the randomly selected candidate μ_ℓ according to a random walk process by means of a univariate normal distribution:

$$\mu_\ell^* \sim \mathcal{N}(\mu_\ell^{(t-1)}, \sigma_{\mu_p}^2), \tag{7.17}$$

where the proposal variance $\sigma_{\mu_p}^2$ is specified as shown in Table 7.2. The remaining components are set with the same value as at iteration $t-1$, i.e. $\mu_{\ell'}^* = \mu_{\ell'}^{(t-1)}$ for $\ell \neq \ell'$. The acceptance ratio for the new component candidate based on the updating of the component μ_ℓ^*

is:

$$
\begin{aligned}
r_\mu &= \frac{p(\Omega^*|\boldsymbol{Y})q(\boldsymbol{\mu}^{(t-1)}|\boldsymbol{\mu}^*)}{p(\Omega^{(t-1)}|\boldsymbol{Y})q(\boldsymbol{\mu}^*|\boldsymbol{\mu}^{(t-1)})} \\
&= \frac{p(\mu_\ell^*|\boldsymbol{Y},(\Omega\backslash\mu_\ell)^*)\left(\frac{1}{q-1}\right)q(\mu_\ell^{(t-1)}|\mu_\ell^*)}{p(\mu_\ell^{(t-1)}|\boldsymbol{Y},(\Omega\backslash\mu_\ell)^{(t-1)})\left(\frac{1}{q-1}\right)q(\mu_\ell^*|\mu_\ell^{(t-1)})} \\
&= \frac{p(\mu_\ell^*|\boldsymbol{Y},(\Omega\backslash\mu_\ell)^*)f_{\mathcal{N}}\left(\frac{\mu_\ell^{(t-1)}-\mu_\ell^*}{\sigma_{\mu_p}}\right)}{p(\mu_\ell^{(t-1)}|\boldsymbol{Y},(\Omega\backslash\mu_\ell)^{(t-1)})f_{\mathcal{N}}\left(\frac{\mu_\ell^*-\mu_\ell^{(t-1)}}{\sigma_{\mu_p}}\right)} = \frac{p(\mu_\ell^*|\boldsymbol{Y},(\Omega\backslash\mu_\ell)^*)}{p(\mu_\ell^{(t-1)}|\boldsymbol{Y},(\Omega\backslash\mu_\ell)^{(t-1)})},
\end{aligned}
$$

where $f_{\mathcal{N}}(\cdot)$ is the density of the standard normal distribution and $p(\mu_\ell^*|\boldsymbol{Y},(\Omega\backslash\mu_\ell)^*)$ is the full conditional posterior distribution of Ω^* formulated as

$$
p(\mu_\ell^*|\boldsymbol{Y},(\Omega\backslash\mu_\ell)^*) \propto \prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r^*\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\theta_{rjk}^*\right)^{I(y_{ij}=k)}\right]\frac{1}{\sigma_\mu}\exp\left(-\frac{1}{2\sigma_\mu^2}{\mu_\ell^*}^2\right),
$$

where the unknown variance σ_μ^2 is specified as shown in Table 7.1. Thus, the acceptance ratio r_μ is formulated as

$$
r_\mu = \text{LR} \times \exp\left(-\frac{1}{2\sigma_\mu^2}\left({\mu_\ell^*}^2 - {\mu_\ell^{(t-1)}}^2\right)\right), \tag{7.18}
$$

where LR is defined as (7.13).

- **New candidate for $\boldsymbol{\phi}$**: We update a randomly selected element of $\boldsymbol{\phi}$. The component to update is selected randomly. Therefore, the probability for a particular ϕ_ℓ ($\ell \in 2,\ldots,q-1$) of being selected is $1/(q-2)$. The new candidate $\boldsymbol{\phi}^*$ is as a result of drawing an uniform distribution as

$$
\phi_\ell^* \sim \mathcal{U}\left[\phi_{\ell-1}^{(t-1)}, \phi_{\ell+1}^{(t-1)}\right] \tag{7.19}
$$

to set the new value for ϕ_ℓ^* at the iteration t. The remaining components are set with the same value as at iteration $t-1$, i.e. $\phi_{\ell'}^* = \phi_{\ell'}^{(t-1)}$ for $\ell \neq \ell'$. This proposal ensures the new state for $\boldsymbol{\phi}^*$ achieves the monotone constraint in the score parameters. From here, we com-

pute the acceptance ratio r_ϕ for the new component candidate $\boldsymbol{\phi}^*$ as

$$r_\phi = \frac{p(\phi_\ell^*|\boldsymbol{Y}, (\Omega\backslash\phi_\ell)^*)\left(\frac{1}{q-2}\right) q(\phi_\ell^{(t-1)}|\phi_\ell^*)}{p(\phi_\ell^{(t-1)}|\boldsymbol{Y}, (\Omega\backslash\phi_\ell)^{(t-1)})\left(\frac{1}{q-2}\right) q(\phi_\ell^*|\phi_\ell^{(t-1)})}$$
$$= \frac{p(\phi_\ell^*|\boldsymbol{Y}, (\Omega\backslash\phi_\ell)^*)\left(\frac{1}{\phi_{\ell+1}^*-\phi_{\ell-1}^*}\right)}{p(\phi_\ell^{(t-1)}|\boldsymbol{Y}, (\Omega\backslash\phi_\ell)^{(t-1)})\left(\frac{1}{\phi_{\ell+1}^{(t-1)}-\phi_{\ell-1}^{(t-1)}}\right)} = \frac{p(\phi_\ell^*|\boldsymbol{Y}, (\Omega\backslash\phi_\ell)^*)}{p(\phi_\ell^{(t-1)}|\boldsymbol{Y}, (\Omega\backslash\phi_\ell)^{(t-1)})}$$

because $\phi_{\ell+1}^* - \phi_{\ell-1}^* = \phi_{\ell+1}^{(t-1)} - \phi_{\ell-1}^{(t-1)}$. Additionally, the full conditional posterior distribution of $\boldsymbol{\phi}^*$ is defined by using the likelihood density (7.5) and the joint prior distribution (7.8) as

$$p(\phi_\ell^*|\boldsymbol{Y}, (\Omega\backslash\phi_\ell)^*) \propto \prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r^*\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\theta_{rjk}^*\right)^{I(y_{ij}=k)}\right]$$
$$\times \left(\phi_{\ell+1}^* - \phi_\ell^*\right)^{\lambda_{\ell+1}-1}\left(\phi_\ell^* - \phi_{\ell-1}^*\right)^{\lambda_\ell-1}.$$

Thus, the acceptance ratio r_ϕ is formulated as,

$$r_\phi = \text{LR} \times \left(\frac{\phi_{\ell+1}^* - \phi_\ell^*}{\phi_{\ell+1}^* - \phi_\ell^{(t-1)}}\right)^{\lambda_{\ell+1}-1}\left(\frac{\phi_\ell^* - \phi_{\ell-1}^*}{\phi_\ell^{(t-1)} - \phi_{\ell-1}^*}\right)^{\lambda_\ell-1}. \quad (7.20)$$

- **New candidate for $\boldsymbol{\alpha}$**: Each row group effect α_r $(r = 1, \ldots, R)$ is updated separately by a Gaussian random walk proposal. The selection of the component α_r to update is done randomly. In order to update α_r in the iteration t, a disturbance δ_α is drawn from

$$\delta_\alpha \sim \mathcal{N}(0, \sigma_{\alpha_p}^2), \quad (7.21)$$

where the unknown variance $\sigma_{\alpha_p}^2$ is specified as shown in Table 7.2. The disturbance δ_α is added to $\alpha_r^{(t-1)}$. In order to preserve the sum-to-zero constraint in the new candidate $\boldsymbol{\alpha}^*$, δ_α is subtracted from $\alpha_{r'}^{(t-1)}$ where r' is the index of another distinct row group which is

also chosen at random:

$$\alpha_r^* = \alpha_r^{(t-1)} + \delta_\alpha \quad \text{and} \quad \alpha_{r'}^* = \alpha_{r'}^{(t-1)} - \delta_\alpha \quad (r \neq r'). \tag{7.22}$$

The remaining components are set with the same value as at iteration $t-1$, i.e. $\alpha_\ell^* = \alpha_\ell^{(t-1)}$ for $\ell \notin \{r, r'\}$. Given equations (7.21) and (7.22), the distribution of the updated component α_ℓ^* given $\boldsymbol{\alpha}^{(t-1)}$ is:

$$\begin{aligned} \alpha_r^* | \boldsymbol{\alpha}^{(t-1)} &\sim \mathcal{N}(\alpha_r^{(t-1)}, \sigma_{\alpha_p}^2), \\ \alpha_{r'}^* &= \alpha_{r'}^{(t-1)} - (\alpha_r^* - \alpha_r^{(t-1)}). \end{aligned} \tag{7.23}$$

Thus, the proposal function for the new candidate $q(\boldsymbol{\alpha}^* | \boldsymbol{\alpha}^{(t-1)})$ is:

$$\begin{aligned} q(\boldsymbol{\alpha}^* | \boldsymbol{\alpha}^{(t-1)}) &= q(\alpha_r^*, \alpha_{r'}^* | \boldsymbol{\alpha}^{(t-1)}) \\ &= \frac{1}{R} \frac{1}{R-1} f_{\mathcal{N}} \left(\frac{\alpha_r^* - \alpha_r^{(t-1)}}{\sigma_{\alpha_p}} \right) \\ &\quad \times \delta(\alpha_r^* + \alpha_{r'}^* + \sum_{\ell \notin \{r,r'\}}^{R} \alpha_\ell^*) \delta^{R-2}(\boldsymbol{\alpha}_{-\{r,r'\}}^* - \boldsymbol{\alpha}_{-\{r,r'\}}^{(t-1)}), \end{aligned}$$

where $\boldsymbol{\alpha}_{-\{r,r'\}}$ indicates the parameter vector $\boldsymbol{\alpha}$ without the components α_r and $\alpha_{r'}$ and $\delta^d(\cdot)$ is the delta function to apply the sum-to-zero constraint on a d-dimensional vector. The acceptance ratio for the new candidate $\boldsymbol{\alpha}^*$ is:

$$\begin{aligned} r_\alpha &= \frac{p(\alpha_r^*, \alpha_{r'}^* | \boldsymbol{Y}, (\Omega \backslash \{\alpha_r, \alpha_{r'}\})^*) q(\boldsymbol{\alpha}^{(t-1)} | \boldsymbol{\alpha}^*)}{p(\alpha_r^{(t-1)}, \alpha_{r'}^{(t-1)} | \boldsymbol{Y}, (\Omega \backslash \{\alpha_r, \alpha_{r'}\})^{(t-1)}) q(\boldsymbol{\alpha}^* | \boldsymbol{\alpha}^{(t-1)})} \\ &= \frac{p(\alpha_r^*, \alpha_{r'}^* | \boldsymbol{Y}, (\Omega \backslash \{\alpha_r, \alpha_{r'}\})^*) f_{\mathcal{N}} \left(\frac{\alpha_r^{(t-1)} - \alpha_r^*}{\sigma_{\alpha_p}} \right)}{p(\alpha_r^{(t-1)}, \alpha_{r'}^{(t-1)} | \boldsymbol{Y}, (\Omega \backslash \{\alpha_r, \alpha_{r'}\})^{(t-1)}) f_{\mathcal{N}} \left(\frac{\alpha_r^* - \alpha_r^{(t-1)}}{\sigma_{\alpha_p}} \right)} \\ &= \frac{p(\alpha_r^*, \alpha_{r'}^* | \boldsymbol{Y}, (\Omega \backslash \{\alpha_r, \alpha_{r'}\})^*)}{p(\alpha_r^{(t-1)}, \alpha_{r'}^{(t-1)} | \boldsymbol{Y}, (\Omega \backslash \{\alpha_r, \alpha_{r'}\})^{(t-1)})}. \end{aligned}$$

As indicated above, the prior distribution for $\boldsymbol{\alpha}$ follows the one-dimensional degenerate normal (7.14). Thus, the full conditional

posterior distribution $p(\alpha_r^*|\boldsymbol{Y}, (\Omega\backslash\alpha_r)^*)$ is formulated as:

$$p(\alpha_r^*, \alpha_{r'}^*|\boldsymbol{Y}, (\Omega\backslash\{\alpha_r, \alpha_{r'}\})^*) \propto \prod_{i=1}^{n}\left[\sum_{r=1}^{R}\pi_r^*\prod_{j=1}^{m}\prod_{k=1}^{q}\left(\theta_{rjk}^*\right)^{I(y_{ij}=k)}\right] \times \exp\left(-\frac{1}{2\sigma_\alpha^2}(\alpha_r^{*^2} + \alpha_{r'}^{*^2} + \sum_{\ell\notin\{r,r'\}}^{R}\alpha_\ell^{*^2})\right), \tag{7.24}$$

where the unknown variance σ_α^2 is specified as shown in Table 7.1. Thus, the acceptance ratio r_α is,

$$r_\alpha = \text{LR} \times \exp\left(-\frac{1}{2\sigma_\alpha^2}(\alpha_r^{*^2} + \alpha_{r'}^{*^2} - \alpha_r^{(t-1)^2} - \alpha_{r'}^{(t-1)^2})\right). \tag{7.25}$$

- **New candidate for $\boldsymbol{\beta}$**: Each column effect β_j ($j = 1, \ldots, m$) is updated separately by a Gaussian random walk proposal similar to the previous update in $\boldsymbol{\alpha}$. Thus, the acceptance ratio r_β is formulated as,

$$r_\beta = \text{LR} \times \exp\left(-\frac{1}{2\sigma_\beta^2}(\beta_j^{*^2} + \beta_{j'}^{*^2} - \beta_j^{(t-1)^2} - \beta_{j'}^{(t-1)^2})\right), \tag{7.26}$$

  where LR is defined as (7.13) and σ_β^2 is specified in Table 7.1.

- **New candidate for $\boldsymbol{\gamma}$**: Each interaction factor parameter γ_{rj} ($r = 1, \ldots, R$ and $j = 1, \ldots, m$) is updated separately by using a random walk Metropolis proposal. Again we must take care to preserve the zero means in each row and column of the interaction matrix $\boldsymbol{\gamma}^*$ in the new candidate. Therefore, a disturbance δ_γ is drawn from a univariate normal distribution:

$$\delta_\gamma \sim \mathcal{N}(0, \sigma_{\gamma_p}^2),$$

  where the proposal variance $\sigma_{\gamma_p}^2$ is specified as shown in Table 7.2. In order to generate the new candidate $\boldsymbol{\gamma}^*$, we proceed in the following way: First, the disturbance δ_γ is added to a randomly selected component of the iteration matrix γ_{uv}. Then, we randomly

select another row u' and column v' which the disturbance δ_γ is subtracted and added in the following way:

$$\begin{aligned}
&\gamma^*_{uv} = \gamma^{(t-1)}_{uv} + \delta_\gamma \quad \gamma^*_{uv'} = \gamma^{(t-1)}_{uv'} - \delta_\gamma \\
&\gamma^*_{u'v} = \gamma^{(t-1)}_{u'v} - \delta_\gamma \quad \gamma^*_{u'v'} = \gamma^{(t-1)}_{u'v'} + \delta_\gamma \quad (u \neq u', v \neq v').
\end{aligned}$$

The remaining components are set with the same value as at iteration $t-1$, i.e. $\gamma^*_{\ell\vartheta} = \gamma^{(t-1)}_{\ell\vartheta}$ for $\ell \notin \{u, u'\}$ and $\vartheta \notin \{v, v'\}$. Furthermore, the distribution of the updated component $\gamma^*_{\ell\vartheta}$ given $\boldsymbol{\gamma}^{(t-1)}$ is:

$$\gamma^*_{uv}|\boldsymbol{\gamma}^{(t-1)} \sim \mathcal{N}(\gamma^{(t-1)}_{uv}, \sigma^2_{\gamma_p}). \tag{7.27}$$

The probability of selecting a random component γ_{uv} from the interaction matrix is $\frac{1}{Rm}$ and therefore the probabilities of selecting the corresponding elements in row u' and column v' are $\frac{1}{R-1}$ and $\frac{1}{m-1}$ respectively. Thus, the proposal function for the new candidate $q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t-1)})$ is:

$$\begin{aligned}
q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t-1)}) &= q(\gamma^*_{uv}|\boldsymbol{\gamma}^{(t-1)}) \\
&= \frac{1}{Rm}\frac{1}{R-1}\frac{1}{m-1} f_{\mathcal{N}}\left(\frac{\gamma^*_{uv} - \gamma^{(t-1)}_{uv}}{\sigma_{\gamma_p}}\right) \\
&\quad \times \delta(\gamma^*_{uv} + \gamma^*_{u'v} + \gamma^*_{uv'} + \gamma^*_{u'v'} + \sum_{\ell \notin \{u,u'\}}^{R} \sum_{\vartheta \notin \{v,v'\}}^{m} \gamma^*_{\ell\vartheta}) \\
&\quad \times \delta^{Rm-4}(\boldsymbol{\gamma}^*_{-UV} - \boldsymbol{\gamma}^{(t-1)}_{-UV}).
\end{aligned}$$

where $\boldsymbol{\gamma}_{-UV}$ indicates the parameter matrix $\boldsymbol{\gamma}$ without the components γ_{uv}, $\gamma_{u'v}$, $\gamma_{uv'}$, and $\gamma_{u'v'}$. The acceptance ratio for the new candidate $\boldsymbol{\gamma}^*$ is:

$$\begin{aligned}
r_\gamma &= \frac{p(\boldsymbol{\gamma}^*_{UV}|\boldsymbol{Y}, (\Omega\backslash\boldsymbol{\gamma}_{UV})^*) q(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{\gamma}^*)}{p(\boldsymbol{\gamma}^{(t-1)}_{UV}|\boldsymbol{Y}, (\Omega\backslash\boldsymbol{\gamma}_{UV})^{(t-1)}) q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t-1)})} \\
&= \frac{p(\boldsymbol{\gamma}^*_{UV}|\boldsymbol{Y}, (\Omega\backslash\boldsymbol{\gamma}_{UV})^*) f_{\mathcal{N}}\left(\frac{\gamma^{(t-1)}_{uv} - \gamma^*_{uv}}{\sigma_{\gamma_p}}\right)}{p(\boldsymbol{\gamma}^{(t-1)}_{UV}|\boldsymbol{Y}, (\Omega\backslash\boldsymbol{\gamma}_{UV})^{(t-1)}) f_{\mathcal{N}}\left(\frac{\gamma^*_{uv} - \gamma^{(t-1)}_{uv}}{\sigma_{\gamma_p}}\right)} = \frac{p(\boldsymbol{\gamma}^*_{UV}|\boldsymbol{Y}, (\Omega\backslash\boldsymbol{\gamma}_{UV})^*)}{p(\boldsymbol{\gamma}^{(t-1)}_{UV}|\boldsymbol{Y}, (\Omega\backslash\boldsymbol{\gamma}_{UV})^{(t-1)})},
\end{aligned}$$

where $\boldsymbol{\gamma}_{UV} = \{\gamma_{uv}, \gamma_{u'v}, \gamma_{uv'}, \gamma_{u'v'}\}$ and $\Omega\backslash\boldsymbol{\gamma}_{UV}$ indicates the param-

eter vector Ω without the parameter set $\boldsymbol{\gamma}_{UV}$. The prior distribution for $\boldsymbol{\gamma}$ follows the two-dimensional degenerate normal (7.16). Similarly to the update in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ update, the acceptance ratio r_γ is

$$r_\gamma = \text{LR} \times \exp\left\{-\frac{1}{2\sigma_\gamma^2}\left(\sum_{\ell\in\{u,u'\}}^{R}\sum_{\vartheta\in\{v,v'\}}^{m}\gamma_{\ell\vartheta}^{*^2} - \sum_{\ell\in\{u,u'\}}^{R}\sum_{\vartheta\in\{v,v'\}}^{m}\gamma_{\ell\vartheta}^{(t-1)^2}\right)\right\}, \tag{7.28}$$

where σ_γ^2 is specified as shown in Table 7.1.

- **New candidate for $\boldsymbol{\pi}$**: Each row membership probability π_r ($r = 1,\ldots,R$) is updated separately by exchanging probability between two row groups. We randomly select a pair of candidate clusters to update, ℓ and ℓ', and calculate the proportion of their combined probability assigned to group ℓ at iteration $t-1$:

$$\omega_\pi = \frac{\pi_\ell^{(t-1)}}{\pi_\ell^{(t-1)} + \pi_{\ell'}^{(t-1)}}.$$

We draw a new proportion ω_π' from a logistic normal distribution with mean ω_π and constant variance $\sigma_{\pi_p}^2 = 0.3$, which is defined as:

$$f_{\text{logistic}\mathcal{N}}(\omega_\pi' \mid \boldsymbol{\pi}^{(t-1)}) = f_{\mathcal{N}}(\text{logit}(\omega_\pi'))\frac{1}{\omega_\pi'(1-\omega_\pi')}, \tag{7.29}$$

where $f_{\mathcal{N}}(\cdot)$ is the pdf of a normal distribution with mean $\text{logit}(\omega_\pi)$ and variance $\sigma_{\pi_p}^2$. As the new candidate $\boldsymbol{\pi}^*$ is only valid if the sum-to-one constraint is preserved, we update the two random candidates ℓ and ℓ' as follows:

$$\pi_\ell^* = \omega_\pi'(\pi_\ell^{(t-1)} + \pi_{\ell'}^{(t-1)}) \quad \text{and} \quad \pi_{\ell'}^* = (1-\omega_\pi')(\pi_\ell^{(t-1)} + \pi_{\ell'}^{(t-1)}).$$

The remaining components are set with the same value as at iteration $t-1$, i.e. $\pi_s^* = \pi_s^{(t-1)}$ for $s \notin \{\ell, \ell'\}$.
The acceptance ratio for the new candidate $\boldsymbol{\pi}^*$ is:

$$r_\pi = \frac{p(\pi_\ell^*|\boldsymbol{Y}, (\Omega\backslash\pi_\ell)^*)\left(\frac{1}{R}\right)\left(\frac{1}{R-1}\right)q(\pi_\ell^{(t-1)}|\boldsymbol{\pi}^*)}{p(\pi_\ell^{(t-1)}|\boldsymbol{Y}, (\Omega\backslash\pi_\ell)^{(t-1)})\left(\frac{1}{R}\right)\left(\frac{1}{R-1}\right)q(\pi_\ell^*|\boldsymbol{\pi}^{(t-1)})}. \tag{7.30}$$

In order to compute $q(\boldsymbol{\pi}^* \mid \boldsymbol{\pi}^{(t-1)})$, we apply the variable change

$$\begin{aligned} q(\boldsymbol{\pi}^* \mid \boldsymbol{\pi}^{(t-1)}) = q(\pi_\ell^* \mid \boldsymbol{\pi}^{(t-1)}) &= f_{\text{logistic}\mathcal{N}}(\omega'_\pi \mid \boldsymbol{\pi}^{(t-1)}) \left| \frac{\partial \omega'_\pi}{\partial \pi_\ell^*} \right| \\ &= f_{\text{logistic}\mathcal{N}}(\omega'_\pi \mid \boldsymbol{\pi}^{(t-1)}) \frac{1}{\pi_\ell^{(t-1)} + \pi_{\ell'}^{(t-1)}}, \end{aligned} \tag{7.31}$$

where $f_{\text{logistic}\mathcal{N}}(\cdot|\boldsymbol{\pi}^{(t-1)})$ is the pdf formulated in (7.29). Finally, using eq. (7.31) and (7.29) the proposal distribution ratio is obtained:

$$\begin{aligned} \frac{q(\pi_\ell^{(t-1)} \mid \boldsymbol{\pi}^*)}{q(\pi_\ell^* \mid \boldsymbol{\pi}^{(t-1)})} &= \frac{f_{\mathcal{N}}(\text{logit}(\omega'_\pi))\omega'_\pi(1-\omega'_\pi)(\pi_\ell^{(t-1)} + \pi_{\ell'}^{(t-1)})}{f_{\mathcal{N}}(\text{logit}(\omega'_\pi))\omega_\pi(1-\omega_\pi)(\pi_\ell^* + \pi_{\ell'}^*)} \\ &= \frac{\omega'_\pi(1-\omega'_\pi)}{\omega_\pi(1-\omega_\pi)} = \frac{\pi_\ell^* \pi_{\ell'}^*}{\pi_\ell^{(t-1)} \pi_{\ell'}^{(t-1)}}, \end{aligned}$$

because $\pi_\ell^* + \pi_{\ell'}^* = \pi_\ell^{(t-1)} + \pi_{\ell'}^{(t-1)}$. Additionally, the marginal posterior distribution of $\boldsymbol{\pi}$ at iteration t is defined by using the likelihood density (7.5) and the prior Dirichlet distribution with hyperparameter vector $\boldsymbol{\lambda}_\pi$ (see Table 7.1) as:

$$p(\pi_\ell^*|\boldsymbol{Y}, (\Omega \backslash \pi_\ell)^*) \propto \prod_{i=1}^{n} \left[\sum_{r=1}^{R} \pi_r^* \prod_{j=1}^{m} \prod_{k=1}^{q} \left(\theta_{rjk}^*\right)^{I(y_{ij}=k)} \right] (\pi_\ell^*)^{\lambda_\ell - 1} (\pi_{\ell'}^*)^{\lambda_{\ell'} - 1}.$$

Thus, the acceptance ratio r_π is formulated as,

$$r_\pi = \text{LR} \times \left(\frac{\pi_\ell^*}{\pi_\ell^{(t-1)}} \right)^{\lambda_\ell - 1} \left(\frac{\pi_{\ell'}^*}{\pi_{\ell'}^{(t-1)}} \right)^{\lambda_{\ell'} - 1}. \tag{7.32}$$

(c) We accept the new candidate $\Omega^{(t)} = \Omega^*$ if a random draw from an $\mathcal{U}(0,1)$ distribution is less than the minimum between 1 and the acceptance ratio. Otherwise, we reject it and set $\Omega^{(t)} = \Omega^{(t-1)}$.

(d) Change the iteration from t to $t+1$.

3. Return the values $\{\Omega^{(B+1)}, \ldots, \Omega^{(T)}\}$, where B is the number of iterations used in the burn-in period .

4. Test whether convergence has been achieved. If not, increase T and return to step 2.

Summary details of the default parameter settings are collected together in Table 7.2. We use Geweke Time Series, ESS and Heidelberger and Welch diagnostics to test the convergence of a single chain and the Gelman and Rubin diagnostic to test the convergence using several chains. In the latter, we pool all the MCMC chains used once the convergence is reached (see brief description in Appendix F.2).

Table 7.2: Default settings for the parameters controlling the proposal distributions for the Metropolis-Hastings sampler in the estimation of the row clustering model.

Move Parameter	Proposal Constants
$\{\mu_k\}$	$\sigma^2_{\mu_p} = 0.3$
$\{\alpha_r\}$	$\sigma^2_{\alpha_p} = 0.3$
$\{\beta_j\}$	$\sigma^2_{\beta_p} = 0.3$
$\{\gamma_{rj}\}$	$\sigma^2_{\gamma_p} = 0.3$
$\{\pi_r\}$	$\sigma^2_{\pi_p} = 0.3$

7.2.5 Label Switching Problem

One of the most common drawbacks associated with the MCMC application of mixture models is the non-identifiability of the labels of the mixture components. For example, the two fitted row cluster mixture models $\widehat{\pi}_1 f(\widehat{\Theta}_1; \boldsymbol{Y}) + \widehat{\pi}_2 f(\widehat{\Theta}_2; \boldsymbol{Y})$ and $\widehat{\pi}_2 f(\widehat{\Theta}_2; \boldsymbol{Y}) + \widehat{\pi}_1 f(\widehat{\Theta}_1; \boldsymbol{Y})$ have the same likelihood. Therefore, we cannot uniquely identify $\widehat{\pi}_1 f(\widehat{\Theta}_1; \boldsymbol{Y})$ as the "first" component of the mixture. The components may be ordered arbitrarily. As a consequence, various functions of interest such as the marginal posterior distribution of the parameters and their associated moments may be invariant under permutations of the labels of its components. In a Bayesian inference approach, if the prior distributions are the same for all the permutations of parameters (i.e. they do not distinguish between the mixture components) then the posterior distributions will show modes which will be similarly symmetric (Stephens, 2000a). This difficulty is so-called *label switching* problem (see for example Stephens (2000a), Jasra et al. (2005), and Marin and Robert (2007, Section 6.4) for a review and illustrative examples of this problem). In our clustering model approach, this problem is crucial as clustering inference requires an unequivocal assignment of the labels to the mixture components. Therefore, a procedure addressing the label switching problem is required to both reach con-

vergence of a MCMC sampler and produce a satisfactory Bayesian analysis of the data. Figure 7.1 illustrates the effects of label switching on a simulated data when fitting a row clustering stereotype model ($\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$) by using a Metropolis-Hastings sampler, as introduced in Section 7.2.4. Multimodality in the estimated marginal posterior densities and distinct jumps in the traces of row effects $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are indicative of a label switching problem. Obtaining summary statistics from the cluster dependent on the basis of this MCMC output is not straightforward.

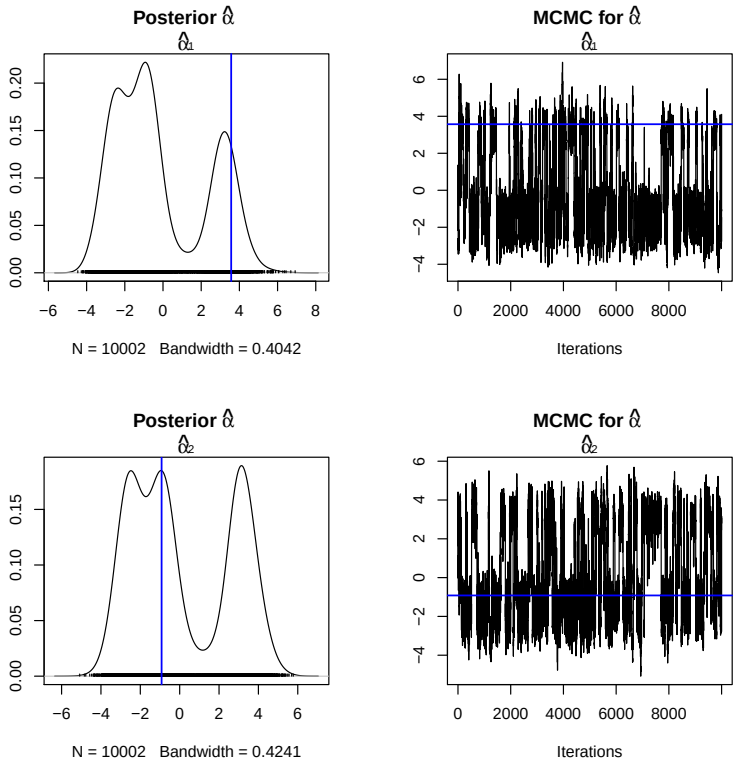

Figure 7.1: ***Label switching**: Marginal posterior densities and trace plots of row effects $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ when fitting a row clustering stereotype model ($\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$). The sample size is $n = 500$, the number of categories is $q = 4$, the number of columns is $m = 3$ and the number of row clusters is $R = 3$. The fitting is performed with the Metropolis-Hastings sampler (Section 7.2.4). The blue line is the true parameter value. Jumps in the trace plots and multimodality in the densities indicates that label switching problem is occurring.*

There have been several possible strategies proposed to deal with the label switching problem. A good review of them can be found in Stephens (2000a, Sec-

tion 3), McLachlan and Peel (2000, Section 4.9), Jasra et al. (2005, Section 1.3), and Sperrin et al. (2010). A common solution is to impose an *identifiability constraint* (IC) on the parameter space such as $\alpha_1 < \alpha_2 < \ldots < \alpha_R$ with the aim of obtaining marginal posterior distributions satisfying this constraint. This constraint is chosen so that just one labelling permutation satisfies the constraint in each iteration. This solution is simple and works well in many situations. However, many choices of IC will be ineffective to solve the problem because chain mixing will be impeded (see Stephens (2000a, Section 3.1) for a descriptive example). Additionally, McLachlan and Peel (2000, Section 4.9) observed that estimations of adjacent parameters will be a biased sample (over-estimated) if we sample under the ordering constraint. This is due to the fact that the IC is providing information which limits the prior distribution of the parameters despite the fact that the prior distributions were non-informative in relation to the components of the mixture model. Even if an appropriate prior is specified, one which incorporates the IC, MCMC mixing may be slowed or made difficult by the constraint.

Other approaches avoiding the identifiability constraint difficulties have been proposed. Most of them are algorithms based on deterministic posterior relabelling of the MCMC outputs. For example, Stephens (2000a) and Celeux (1998) described *relabelling algorithms* based on the k-means clustering of the MCMC outputs with the purpose of obtaining unimodal marginal posterior distributions. Moreover, the *pivotal reordering* strategy described in Marin and Robert (2007, Section 6.4) which consists of selecting one of the modes of the posterior distribution and relabel the MCMC output according to the vicinity to that mode. Additionally, Frühwirth-Schnatter (2001) developed a *random permutation sampler* designed to improve the mixing in a MCMC output when an IC is applied. Celeux et al. (2000) and Hurn et al. (2003) used a decision theoretic approach to minimize the posterior expectation of *label invariant loss functions*.

In this thesis, the relabelling algorithm described in Stephens (2000a) was implemented. Its outline is described in Appendix G. Technical details can be found in the Section 4.1 from his paper. In brief, the procedure is based on measuring the loss for reporting the membership distribution matrix $\widehat{Z} = (\widehat{z}_{ir})$ where $\widehat{z}_{ir}$ is the mean over all the samples of the probability of observation i ($i = 1, \ldots, n$) being classified a posteriori into cluster r ($r = 1, \ldots, R$). The $n \times R$ matrix $\widehat{Z}$ represents the estimated distribution on R-group clusters of the data and the loss for reporting this matrix instead of the true distribution on clustering $P = (p_{ir})$ is

measured by using the Kullback-Leibler divergence ($\mathrm{D_{KL}}$, see its general definition in eq. (3.2) on page 39) in the case of discrete distributions. Thus,

$$\mathrm{D_{KL}}(P, \widehat{Z}) = \sum_{i=1}^{n} \sum_{r=1}^{R} p_{ir} \log \left(\frac{p_{ir}}{\widehat{z}_{ir}} \right).$$

For our approach, $p_{ir} = P\left[z_{ir} = 1 \mid \{y_{ij}\}, \Omega\right]$ which is described in eq. (2.18) on page 30. The procedure finds the minimum p_{ir} of $\mathrm{D_{KL}}(P, \widehat{Z})$ under all permutations of the parameter vectors that are involved in the label switching problem. This $\mathrm{D_{KL}}$ measure can be seen as the expectation of the logarithmic difference between the distributions P and $\widehat{Z}$, where the expectation is taken using the probabilities $\{p_{ir}\}$. It can also be formulated as the difference of two entropies:

$$\mathrm{D_{KL}}(P, \widehat{Z}) = -\sum_{i=1}^{n} \sum_{r=1}^{R} p_{ir} \log(\widehat{z}_{ir}) + \sum_{i=1}^{n} \sum_{r=1}^{R} p_{ir} \log(p_{ir}),$$

where the first term is the cross entropy between distributions P and $\widehat{Z}$ and the second term is the entropy of P.

Figure 7.2 shows the marginal posterior distributions and the trace plots for the row effect parameters α_1 and α_2 after applying this relabelling procedure to the MCMC samples shown in Figure 7.1. The results are very satisfactory as the multimodality in the density plots has been removed and the trace plots show good mixing. Thus, the algorithm was effectively able to solve the label switching problem for the parameters. To conclude, in order to obtain a satisfactory Bayesian analysis of the data and reach convergence for our finite mixture approach, the combination of using a MCMC sampler and, later, a label switching procedure is a required strategy.

7.2.6 Simulation Study. One-Dimensional Clustering

We set up a simulation study to test how reliably we were able to estimate the parameters of our one-dimensional clustering approach using the Metropolis-Hastings sampler developed in Section 7.2.4. Similarly to the simulation study for testing the fitting by using the EM algorithm (Section 4.1), the experiment consisted of the simulation of datasets and then fit the correct model to those data.

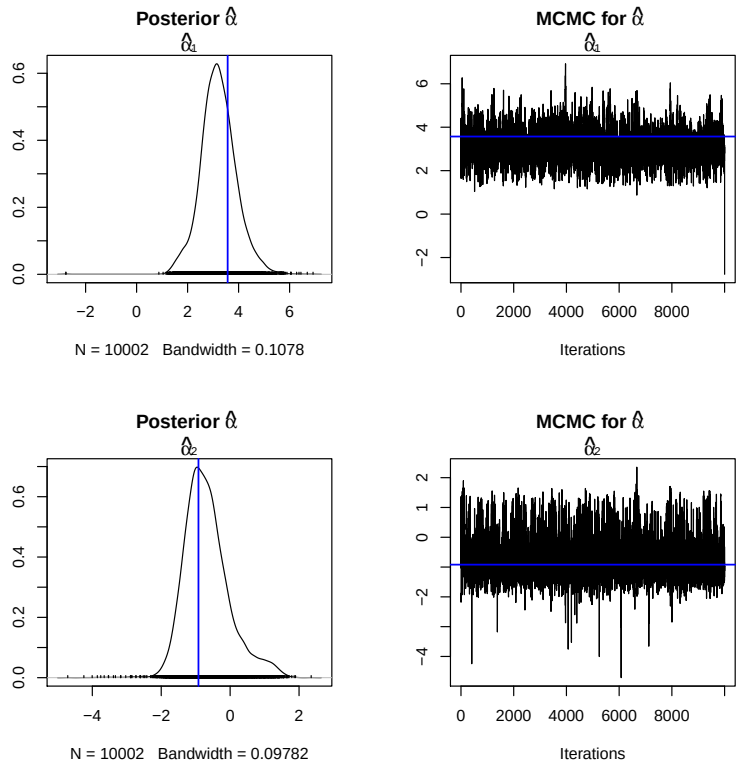

Figure 7.2: ***Label switching***: *Marginal posterior densities and trace plots of row effects* $\widehat{\alpha}_1$ *and* $\widehat{\alpha}_2$ *when fitting a row clustering stereotype model* ($\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$). *The sample size is* $n = 500$, *the number of categories is* $q = 4$, *the number of columns is* $m = 3$ *and the number of row clusters is* $R = 3$. *The fitting is performed with the Metropolis-Hastings sampler (Section 7.2.4). The blue line is the true parameter value. Unimodality in the densities and a satisfactory mixing in the trace plots indicates that the label switching problem is solved.*

The design of the study includes an ordinal response variable with $q = 4$ categories, sample size $n = 500$, number of columns ($m = 3, 5$), and we varied the number of row clusters ($R = 2, 3, 4$). For each combination of number of row clusters, a single set of parameters values was chosen and $H = 100$ data sets (replicates) were generated from an underlying row clustering model. For each data set, we assessed the convergence of the M-H sampler by running $S = 3$ chains in parallel from widely dispersed starting points. We ran each chain for a initial 20000 iterations which were discarded (burn-in period). We then ran each chain for a further 100000 iterations, storing only every 5th state (thinning period). We used the convergence diagnostics described in Section 7.1.3 and Appendix F to

asses the convergence of these chains. This simulation study procedure is outlined in Appendix H.1.

We have simulated over several scenarios. As we are interested in testing the success of parameter estimation in challenging situations where it might be expected that estimation might be difficult, we chose two particular scenarios. The first case is when two of the score parameters $\{\phi_k\}$ have equal values and therefore we could in fact merge their corresponding response categories. A second scenario is to set a very small *a priori* membership probability (e.g. $\pi_2 = 0.015$) and, consequently, few observations will be seen and classified as members of that cluster.

For each chain and replicate, we summarised results computing mean, median, standard deviation, time series standard error, 95% highest posterior density interval (HPD) and Gelman-Rubin's potential scale reduction factor (PSRF) for the elements of the free parameter vector $\Omega = (\{\mu_k\}, \{\phi_k\}, \{\alpha_r\}, \{\beta_j\}, \{\pi_r\})$. The average over the $HS = 300$ chains for each statistical measure are shown in Tables 7.3 and 7.4. The results show that the mean and median of all the parameters are close to their true values and as expected the 95% HPD credible intervals include the true parameters in all the cases. Additionally, Gelman-Rubin's PSRF values are less than 1.2 throughout all the scenarios diagnosing that convergence was reached. For the specific case of the 2 particular scenarios shown in Table 7.4, these results are very satisfactory because our M-H sampler can identify these particular scenarios and get back values close to the true parameters. The marginal posterior distributions, trace plots for each parameter and an illustration of convergence diagnostic plots related to the all the scenarios of this simulation study are shown in Appendix H.2.

7.2.7 Real-Life Data Examples Using M-H Sampler

In this section, we use our Metropolis-Hasting sampler to estimate the parameters for two real-life data examples: the Applied Statistics feedback forms and the tree presences in great smoky mountains. The description for both examples was given in Sections 4.2.1 and 4.2.2 respectively. In the first example, we ran 6 different samplers fitting the row clustering model. Equivalently, we fitted the row clustering including iteration factors model in the second example.

In both examples, we assessed the convergence of the each M-H sampler by

Table 7.3: **MH simulation study**: Summary statistics for estimated parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The true parameter values, Gelman-Rubin's PSRF, mean, median, standard deviation (SD), time-series standard error, and 95% highest posterior density interval (HPD) for different number of row clusters ($R = 2, 3, 4$) are shown. The sample size is $n = 500$, the number of categories is $q = 4$, and the number of columns is $m = 3$.

R	True parameters	Metropolis-Hastings						
		Gelman-Rubin PSRF	Mean	Median	S.D.	Time-series S.E.	HPD 95% lower	HPD 95% upper
2	$\mu_2 = 0.414$	1.0023	0.334	0.547	0.326	0.0080	-0.311	0.965
	$\mu_3 = 2.551$		2.230	2.188	0.361	0.0115	1.582	3.055
	$\mu_4 = 1.507$		1.314	1.274	0.407	0.0128	0.565	2.129
	$\phi_2 = 0.355$		0.246	0.247	0.076	0.0014	0.099	0.396
	$\phi_3 = 0.672$		0.694	0.694	0.041	0.0010	0.616	0.774
	$\alpha_1 = 3.571$		3.185	3.150	0.334	0.0100	2.586	4.013
	$\beta_1 = -0.427$		-0.338	-0.338	0.138	0.0010	-0.615	-0.074
	$\beta_2 = 1.871$		1.785	1.784	0.157	0.0010	1.477	2.092
	$\pi_1 = 0.350$		0.354	0.353	0.041	0.0003	0.275	0.434
3	$\mu_2 = 0.414$	1.0016	0.375	0.347	0.316	0.0103	-0.198	1.019
	$\mu_3 = 2.551$		2.487	2.448	0.432	0.0181	1.729	3.396
	$\mu_4 = 1.507$		1.391	1.349	0.613	0.0249	0.288	2.674
	$\phi_2 = 0.355$		0.284	0.287	0.095	0.0014	0.093	0.464
	$\phi_3 = 0.672$		0.636	0.636	0.056	0.0011	0.524	0.743
	$\alpha_1 = 3.571$		3.201	3.169	0.675	0.0091	1.876	4.637
	$\alpha_2 = -0.919$		-0.698	-0.777	0.583	0.0100	-1.816	0.727
	$\beta_1 = -0.427$		-0.294	-0.294	0.142	0.0011	-0.579	-0.019
	$\beta_2 = 1.871$		1.557	1.552	0.158	0.0019	1.215	1.897
	$\pi_1 = 0.200$		0.230	0.310	0.147	0.0033	0.085	0.629
	$\pi_2 = 0.500$		0.416	0.317	0.140	0.0031	0.208	0.794
4	$\mu_2 = 0.414$	1.0113	0.398	0.380	0.301	0.0108	-0.404	0.891
	$\mu_3 = 2.551$		2.133	2.127	0.429	0.0186	1.303	3.108
	$\mu_4 = 1.507$		1.398	1.409	0.700	0.0302	-0.274	2.246
	$\phi_2 = 0.355$		0.299	0.301	0.073	0.0010	0.157	0.442
	$\phi_3 = 0.672$		0.683	0.583	0.048	0.0007	0.586	0.776
	$\alpha_1 = 3.571$		3.738	3.719	0.759	0.0089	2.269	5.197
	$\alpha_2 = -0.919$		-1.038	-0.976	0.794	0.0122	-2.616	0.306
	$\alpha_3 = 1.228$		1.086	1.034	0.743	0.0080	-0.174	2.536
	$\beta_1 = -0.427$		-0.411	-0.410	0.153	0.0012	-0.709	-0.112
	$\beta_2 = 1.871$		1.744	1.740	0.158	0.0018	1.396	2.092
	$\pi_1 = 0.250$		0.248	0.242	0.108	0.0018	0.048	0.411
	$\pi_2 = 0.320$		0.293	0.299	0.076	0.0014	0.101	0.483
	$\pi_3 = 0.150$		0.206	0.204	0.107	0.0019	0.045	0.451

Table 7.4: **MH simulation study. Special scenarios**: Summary statistics for estimated parameters for stereotype model including row clustering $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$. The true parameter values, Gelman-Rubin's PSRF, mean, median, standard deviation (SD), time-series standard error, and 95% highest posterior density interval (HPD) for $R = 3$ number of row clusters are shown. The sample size is $n = 500$, the number of categories is $q = 4$, and the number of columns is $m = 5$. $\pi_2 = 0.015$ in first subtable and $\phi_2 = \phi_3 = 0.500$ in the second subtable.

R	True parameters	Metropolis-Hastings						
		Gelman-Rubin PSRF	Mean	Median	S.D.	Time-series S.E.	HPD 95% lower	HPD 95% upper
3	$\mu_2 = 0.814$	1.0809	0.827	0.823	0.135	0.0067	0.558	1.093
	$\mu_3 = 0.951$		0.980	0.983	0.182	0.0094	0.616	1.350
	$\mu_4 = 0.207$		0.281	0.294	0.249	0.0133	-0.237	0.801
	$\phi_2 = 0.355$		0.360	0.359	0.019	0.0007	0.323	0.397
	$\phi_3 = 0.672$		0.663	0.664	0.016	0.0004	0.632	0.697
	$\alpha_1 = 3.634$		3.668	3.644	0.335	0.0093	3.528	4.097
	$\alpha_2 = -0.819$		-0.809	-0.816	0.244	0.0116	-1.298	-0.305
	$\beta_1 = -0.427$		-0.547	-0.548	0.114	0.0023	-0.750	-0.311
	$\beta_2 = 1.285$		1.177	1.181	0.110	0.0022	1.012	1.383
	$\beta_3 = 1.872$		2.004	2.003	0.120	0.0028	1.795	2.256
	$\beta_4 = -0.097$		0.037	0.036	0.114	0.0021	-0.189	0.257
	$\pi_1 = 0.400$		0.410	0.411	0.093	0.0024	0.368	0.437
	$\boldsymbol{\pi_2 = 0.015}$		0.053	0.056	0.182	0.0024	0.008	0.089
3	$\mu_2 = 0.814$	1.0798	0.865	0.824	0.237	0.0097	0.466	1.370
	$\mu_3 = 0.951$		1.045	1.002	0.244	0.0104	0.638	1.559
	$\mu_4 = 0.207$		0.276	0.288	0.445	0.0199	-0.271	1.476
	$\boldsymbol{\phi_2 = 0.500}$		0.480	0.480	0.031	0.0006	0.419	0.540
	$\boldsymbol{\phi_3 = 0.500}$		0.499	0.499	0.030	0.0006	0.441	0.559
	$\alpha_1 = 3.634$		3.242	3.271	0.362	0.0100	2.534	3.945
	$\alpha_2 = -0.819$		-0.881	-0.975	0.545	0.0222	-1.729	0.276
	$\beta_1 = -0.427$		-0.498	-0.497	0.148	0.0015	-0.791	-0.206
	$\beta_2 = 1.285$		1.373	1.372	0.155	0.0017	1.082	1.687
	$\beta_3 = 1.872$		1.824	1.822	0.162	0.0018	1.490	2.130
	$\beta_4 = -0.097$		-0.113	-0.154	0.149	0.0016	-0.435	0.051
	$\pi_1 = 0.200$		0.181	0.183	0.034	0.0087	0.111	0.245
	$\pi_2 = 0.500$		0.473	0.461	0.041	0.0070	0.413	0.541

running three chains in parallel from widely dispersed starting points. As in the simulation study (Section 7.2.6), each chain was run for an initial burn-in period of 20000 iterations which are discarded. We then ran each chain for a further 100000 updates, storing only every 5th state (thinning). We used the convergence diagnostics described in Section 7.1.3 and Appendix F to assess the convergence

of these chains. Finally, the results are summarised and compared with those obtained from fitting our suite of models by running the EM algorithm and using AIC to do model comparison (Sections 4.2.1 and 4.2.2).

Example 1: Applied Statistics Course Feedback

The dimensions of the ordinal data matrix for the Applied Statistics course feedback data set are $n = 70$ rows (students) and $m = 10$ columns (questions) where each observation can take one of the three possible categories ($q = 3$). Table 7.5 shows the summary of the MH sampler results for a fitted row clustering model with $R = 3$ clusters. The MLE values from EM algorithm fall within the 95% HPD interval, as expected due to our use of noninformative priors. Moreover, Gelman-Rubin PSRF value diagnoses that the MCMC sampler converged. We graph Figures 7.3-7.4 in order to compare results. Thus, Figure 7.3 depicts the marginal posterior distributions for all the parameters. The expected values of the posterior distribution are very close to the MLE values (blue vertical lines). The trace plots on Figure 7.4 show a good mixing in the sampling of all the parameters.

Example 2: Tree presences in Great Smoky Mountains

The second example relates to the distribution of $n = 41$ tree species along $m = 12$ sites where each observation can take one of the four possible categories ($q = 4$) after categorising the original count data. In order to compare the results with those from the EM algorithm, we fit a row clustering model with interaction factor and $R = 3$ tree (row) groups. A summary of the results and the comparison with the results obtained by using the EM algorithm are shown in Table 7.6. As the previous example, these results are very close to those from EM algorithm. Additionally, the Gelman-Rubin PSRF value diagnoses that the MH sampler converged.

7.3 Discussion

A Bayesian inference procedure for our ordinal stereotype mixture model has been introduced in this chapter. The procedure is based on MCMC sampling

Table 7.5: **Applied Statistics course feedback forms data set:** Estimated parameters for stereotype model including row clustering without interaction factors $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ with $R = 3$ clusters. MLE values from EM algorithm and summary statistics from M-H sampler are shown. The sample size is $n = 70$, the number of categories is $q = 3$, and the number of questions is $m = 10$.

Par.	EM algorithm		Metropolis-Hastings						
	Estim.	S.E.	Gelman-Rubin PSRF	Mean	Median	S.D.	S.E.	HPD 95% lower	HPD 95% upper
$\widehat{\mu}_2$	-0.306	0.140		-0.403	-0.385	0.291	0.0097	-0.964	0.165
$\widehat{\mu}_3$	-2.291	0.307		-2.399	-0.238	0.547	0.0193	-3.524	-1.359
$\widehat{\phi}_2$	0.541	0.195		0.552	0.549	0.066	0.0016	0.426	0.683
$\widehat{\alpha}_1$	3.496	0.346		3.268	3.275	0.564	0.0114	2.112	4.348
$\widehat{\alpha}_2$	-3.571	0.222		-3.390	-3.404	0.276	0.0049	-4.374	-2.434
$\widehat{\beta}_1$	-1.390	0.312		-1.382	-1.373	0.524	0.0086	-2.392	-0.342
$\widehat{\beta}_2$	-2.998	0.351		-2.906	-2.887	0.563	0.0129	-4.202	-1.650
$\widehat{\beta}_3$	-6.272	0.318	1.02124	-5.750	-5.644	0.613	0.0377	-8.491	-3.818
$\widehat{\beta}_4$	0.300	0.437		0.233	0.232	0.442	0.0073	-0.633	1.111
$\widehat{\beta}_5$	1.015	0.432		0.945	0.935	0.436	0.0085	0.082	1.786
$\widehat{\beta}_6$	3.391	0.451		3.255	3.255	0.463	0.0107	2.363	4.145
$\widehat{\beta}_7$	3.561	0.452		3.416	3.421	0.459	0.0098	2.498	4.310
$\widehat{\beta}_8$	3.029	0.463		2.885	2.881	0.489	0.0120	1.930	3.841
$\widehat{\beta}_9$	-1.601	0.332		-1.591	-1.567	0.518	0.0098	-2.697	-0.543
$\widehat{\pi}_1$	0.377	0.218		0.413	0.412	0.107	0.0032	0.200	0.593
$\widehat{\pi}_2$	0.532	0.231		0.437	0.441	0.106	0.0033	0.272	0.671

Notes: The S.E. in Bayesian approach is the Time-series S.E.

using the Metropolis-Hastings algorithm. We have illustrated this Bayesian approach with a simulation study and two real-life data examples. The simulation showed reliable results and the data applications obtained similar results to the inference based on the EM algorithm (Section 2.5). The combination of using the M-H sampler (Sections 7.2.1-7.2.4) and, later, a label switching procedure based on the relabelling algorithm proposed by Stephens (2000a) (Section 7.2.5) is a required strategy in order to obtain a satisfactory Bayesian analysis of the data and reach convergence for our finite mixture approach.

The Bayesian methodology introduced in this chapter does not allow us to make inference regarding the unknown dimension of the model because the M-H algorithm considers a fixed number of components in the mixture. Reversible jump MCMC is however a methodology to estimate the number of clusters and parameters simultaneously. Further development of a RJMCMC sampler to apply to our mixture approach is described in Chapter 8.

Table 7.6: **Tree presences in Great Smoky mountains data set**: Estimated parameters for stereotype model including row clustering with interaction factors $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j} + \boldsymbol{\gamma_{rj}})$ with $R = 3$ clusters. MLE values from EM algorithm and summary statistics from M-H sampler are shown. The sample size is $n = 41$ tree species, the number of categories is $q = 5$, and the number of sites is $m = 12$.

Par.	EM algorithm		Metropolis-Hastings						
	Estim.	S.E.	Gelman-Rubin PSRF	Mean	Median	S.D.	S.E.	HPD 95% lower	HPD 95% upper
$\widehat{\mu}_2$	1.259	0.379		1.684	1.638	0.502	0.0144	0.750	2.684
$\widehat{\mu}_3$	1.828	0.208		1.667	1.652	0.581	0.0169	0.563	2.879
$\widehat{\mu}_4$	0.052	0.268		0.124	0.141	0.868	0.0238	-1.646	1.779
$\widehat{\phi}_2$	0.200	0.197		0.242	0.234	0.098	0.0025	0.055	0.443
$\widehat{\phi}_3$	0.467	0.397		0.423	0.417	0.122	0.0025	0.188	0.657
$\widehat{\phi}_4$	0.987	0.243		0.814	0.837	0.130	0.0025	0.570	0.999
$\widehat{\alpha}_1$	1.083	0.342		1.988	1.919	0.434	0.0241	-0.250	4.331
$\widehat{\alpha}_2$	-5.557	0.378		-4.542	-4.546	0.462	0.0335	-7.330	-1.314
$\widehat{\beta}_1$	3.725	0.238		2.116	2.113	1.990	0.0348	-1.668	6.117
$\widehat{\beta}_2$	1.882	0.233		0.031	0.036	2.067	0.0371	-3.990	4.131
$\widehat{\beta}_3$	1.159	0.372		0.360	0.353	2.040	0.0358	-3.551	4.389
$\widehat{\beta}_4$	-3.311	0.388		-1.449	-1.418	1.847	0.0344	-5.120	2.143
$\widehat{\beta}_5$	-1.440	0.213		-0.689	-0.670	1.833	0.0384	-4.307	2.826
$\widehat{\beta}_6$	-2.286	0.260		-0.727	-0.731	1.795	0.0371	-4.314	2.774
$\widehat{\beta}_7$	-3.134	0.372		-2.306	-2.279	1.969	0.0396	6.308	1.435
$\widehat{\beta}_8$	-2.427	0.287		-0.957	-0.955	1.826	0.0386	-4.659	2.472
$\widehat{\beta}_9$	0.029	0.195		1.370	1.352	1.942	0.0381	-2.435	5.181
$\widehat{\beta}_{10}$	2.584	0.373	1.04140	2.358	2.342	1.963	0.0356	-1.537	6.194
$\widehat{\beta}_{11}$	0.341	0.205		0.532	0.520	2.110	0.0401	-3.625	4.630
$\widehat{\gamma}_{21}$	-1.298	0.350		-0.538	-0.532	1.487	0.0211	-6.361	5.263
$\widehat{\gamma}_{22}$	2.061	0.373		-0.641	-0.661	1.465	0.0299	-6.146	5.412
$\widehat{\gamma}_{23}$	4.189	0.296		0.507	0.478	2.063	0.0351	-5.387	6.606
$\widehat{\gamma}_{24}$	2.861	0.236		0.230	0.232	1.910	0.0206	-5.240	6.064
$\widehat{\gamma}_{25}$	1.748	0.382		0.306	0.291	1.864	0.0283	-5.228	5.921
$\widehat{\gamma}_{26}$	-2.865	0.207		-0.580	0.584	1.902	0.0205	-6.361	4.977
$\widehat{\gamma}_{27}$	2.198	0.188		0.351	0.370	2.295	0.0304	-7.836	9.017
$\widehat{\gamma}_{28}$	2.820	0.374		0.272	0.331	2.278	0.0434	-8.135	8.638
$\widehat{\gamma}_{29}$	3.863	0.328		-1.405	-1.378	2.339	0.0738	-10.663	7.800
$\widehat{\gamma}_{210}$	-1.868	0.307		0.891	0.904	2.269	0.0231	5.633	7.158
$\widehat{\gamma}_{211}$	2.926	0.439		0.891	0.886	2.313	0.0235	-5.648	7.304
$\widehat{\gamma}_{31}$	-2.923	0.312		0.246	0.256	2.281	0.0234	-6.164	6.607
$\widehat{\gamma}_{32}$	3.833	0.382		0.312	0.295	2.414	0.0393	-5.970	6.865
$\widehat{\gamma}_{33}$	2.585	0.309		-1.046	-1.025	2.268	0.0371	-7.735	5.613
$\widehat{\gamma}_{34}$	-0.092	0.297		0.712	0.732	2.229	0.0243	-6.383	7.076
$\widehat{\gamma}_{35}$	-3.228	0.207		0.664	0.693	2.404	0.0263	-6.177	7.081
$\widehat{\gamma}_{36}$	4.311	0.355		-0.570	-0.589	2.313	0.0492	-7.841	6.565
$\widehat{\gamma}_{37}$	2.405	0.393		-0.301	-0.117	2.004	0.0212	-6.023	5.788
$\widehat{\gamma}_{38}$	3.391	0.294		0.800	-0.299	2.030	0.0218	-6.353	5.436
$\widehat{\gamma}_{39}$	3.305	0.259		0.615	0.618	2.161	0.0332	-5.423	6.965
$\widehat{\gamma}_{310}$	1.601	0.293		-0.456	-0.456	2.007	0.0225	-6.314	5.408
$\widehat{\gamma}_{311}$	-3.654	0.387		-0.370	-0.354	2.011	0.0245	-6.071	5.701
$\widehat{\pi}_1$	0.315	0.388		0.338	0.333	0.111	0.0023	0.132	0.559
$\widehat{\pi}_2$	0.277	0.346		0.321	0.314	0.106	0.0022	0.125	0.529

Notes: The S.E. in Bayesian approach is the Time-series S.E.

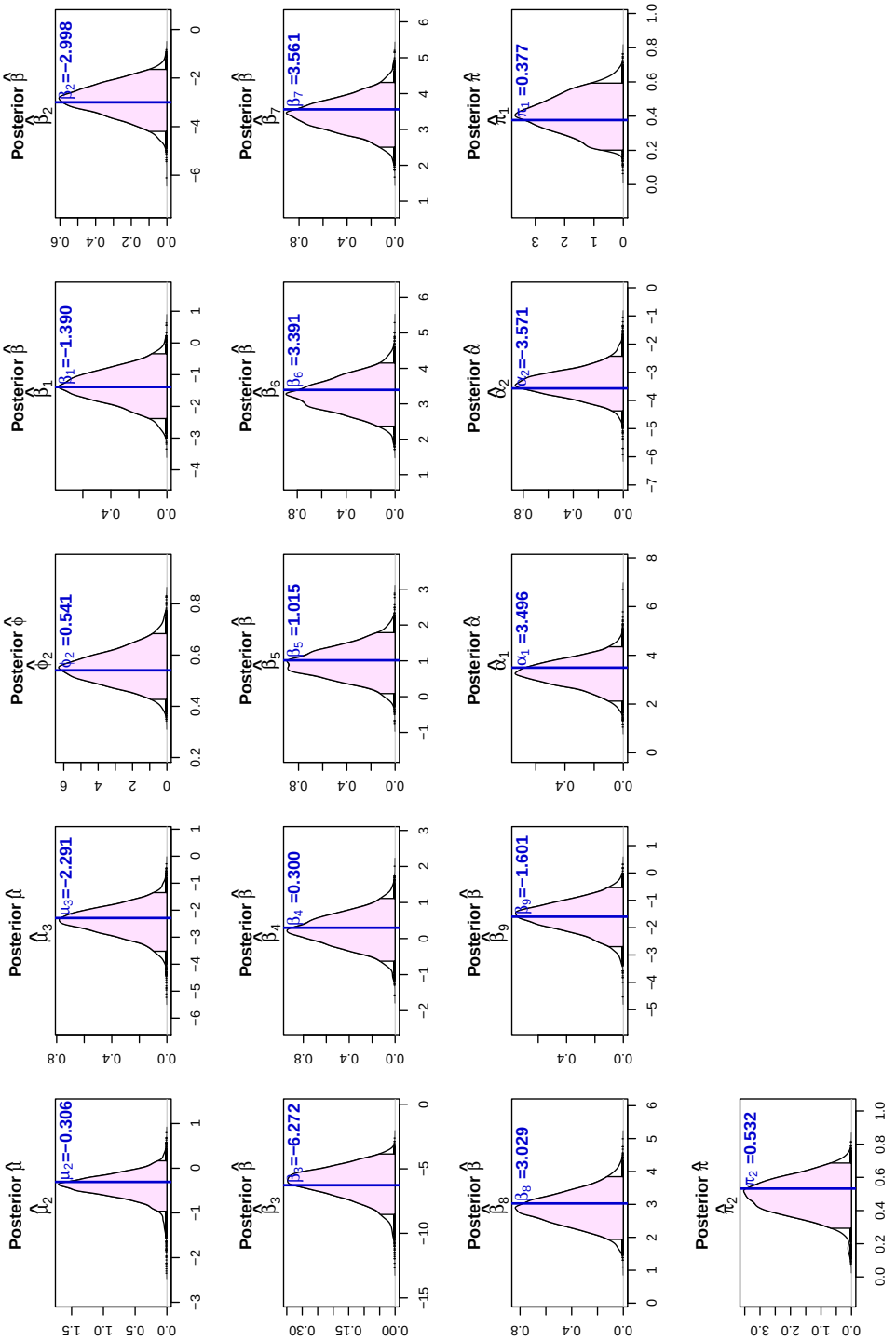

Figure 7.3: ***Applied Statistics course feedback forms data set:*** *Density plot depicting the marginal posterior distribution of the parameters for stereotype model including row clustering without interaction factors model* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ *with* $R = 3$ *row (student) clusters. The blue vertical lines are the MLE values from EM algorithm. 95% HPD credible intervals are shown with shading area.*

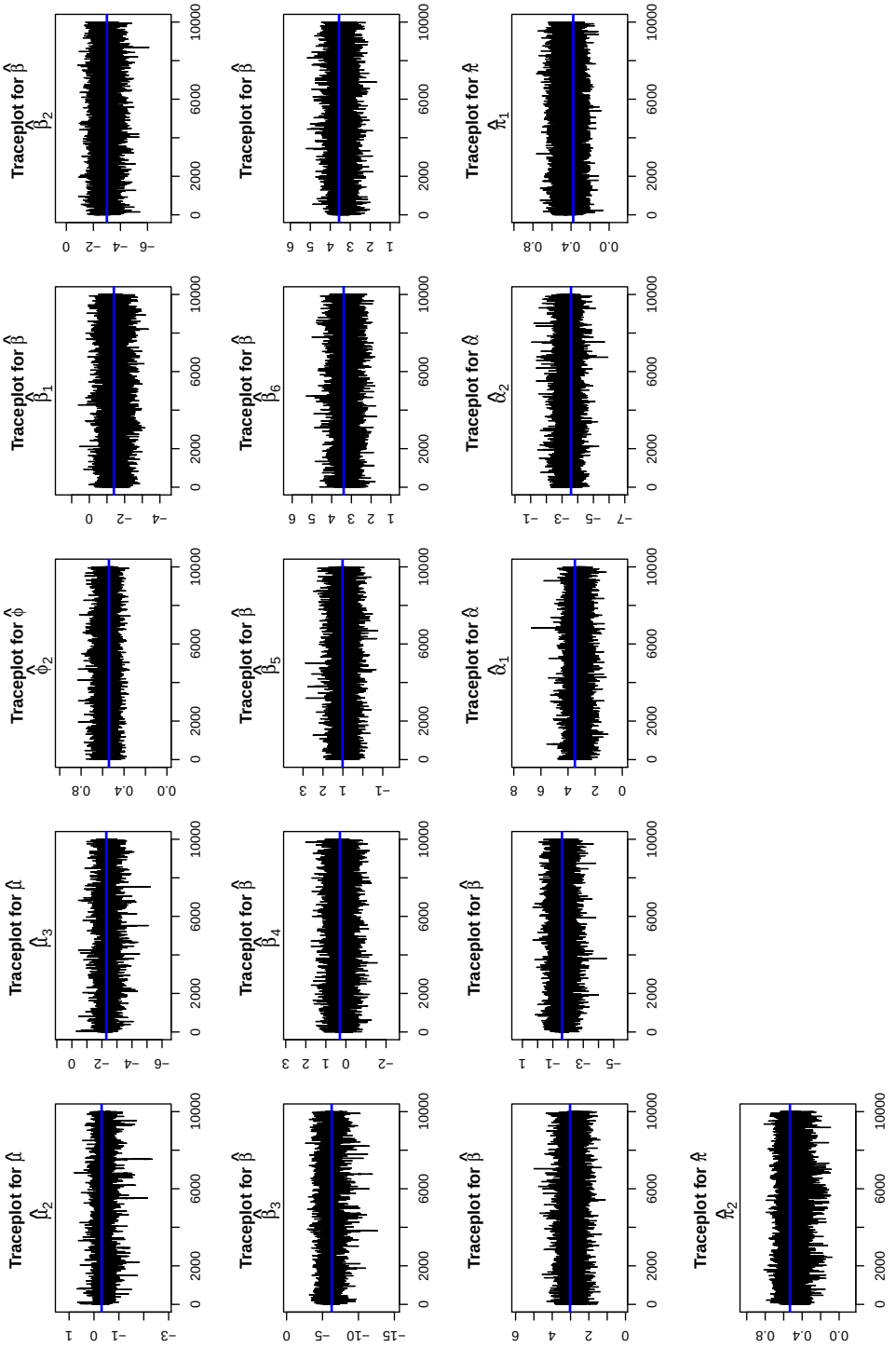

Figure 7.4: ***Applied Statistics course feedback forms data set:*** *Trace plot of the parameters for stereotype model including row clustering without interaction factors model* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ *with* $R = 3$ *row (student) clusters. The blue horizontal lines are the MLE values from EM algorithm.*

Chapter 8

Inference in the Bayesian Paradigm. Variable Dimension

8.1 Introduction. Reversible Jump MCMC Sampler

One of the key difficulties in the use of mixture models to provide a model-based clustering is inference of the unknown dimension of the model, i.e. the number of clusters K most suitable to the data set (see e.g. Böhning and Seidel (2003) and Böhning et al. (2007) for a brief summary). Numerous methods have been proposed to estimate the number of components K in a mixture model. Some research has provided estimates of the marginal likelihoods of K components and used Bayes theorem to obtain the posterior distribution of K (see e.g. Nobile (1994) and Roeder and Wasserman (1997)), or to test K versus $K+1$ number of components (Carlin and Chib (1995); Chib (1995); Raftery (1996)). Mengersen and Robert (1996) introduced a methodology derived from the Kullback-Leibler divergence ($\mathrm{D_{KL}}$, see its general definition in eq. (3.2) on page 39) to compute the distance between mixture models with K and $K+1$ components. Another technique is based on a MCMC sampler using a composite model where the posterior distribution of K can be estimated by the relative frequency with which each model is visited during the simulation (see Phillips and Smith (1996)). Richardson and Green (1997) described a similar sampler based on reversible jump MCMC (RJMCMC) algorithm, which was introduced by Green (1995) and is capable of jumping between parameter subspaces corresponding to different number of components K. An alternative is the birth-and-death process (Stephens, 2000b),

whose mechanism has been shown to be essentially the same as RJMCMC algorithm (Cappé et al., 2003).

RJMCMC is the methodology we have used to estimate the number of clusters and parameters simultaneously from their joint posterior distribution for our clustering mixture approach. In Section 7.2, we introduced the Metropolis-Hastings (M-H) sampler for our approach. RJMCMC can be seen as a trans-dimensional extension of the M-H algorithm. On the one hand, the simple M-H algorithm considers a fixed number of components in the mixture so that it is necessary to apply the algorithm to the data as many times as the number of components we want to estimate. In other words, a run of this algorithm for a particular number of clusters K is independent from another run for another number of groups K' ($K \neq K'$). Once we have the results from all the independent runs, we can apply a model selection procedure based on the deviance information criterion (DIC) (see Section 7.1.4 for details) in order to select the best model, and consequently the appropriate number of clusters. Alternatively, the implementation described in Link and Barker (2010, Chapter 7) and Barker and Link (2013) may be applied. It consists of fitting a relatively small number of models one at a time and then to be compared. In that case, BIC can be used as an approximation to the Bayes factor to compute posterior model weights. On the other hand, the RJMCMC algorithm allows us to simulate the posterior distributions when the dimension of the model is unknown, can vary and has to be estimated. This algorithm allows us to estimate the model parameters and, concurrently, explore the space of models of different dimensions. Therefore, a RJMCMC sampler incorporates the model selection procedure into the estimation stage. Moreover, when the number of clusters is not clearly determined, the RJMCMC approach is a natural way to implement model averaging. Posterior summaries of quantities independent of the number of clusters is simple and straightforward, with models of different dimension appropriately weighted by their posterior probabilities.

In the following sections, the basics of the RJMCMC algorithm are explained (Section 8.2). The use of this algorithm for our approach devising prior distributions, proposal distributions and reversible jump moves for the parameters (including the number of components) is defined in Section 8.3. Assessment of convergence for RJMCMC samplers is described in Section 8.4. A simulation study and two real-life data examples are analysed in Section 8.5 and Section 8.6. Closing remarks are in Section 8.7. Additionally, Appendix I outlines the Castel-

loe and Zimmerman (2002) methodology to assess the convergence of RJMCMC samplers and applies this method to the two real-life data examples.

8.2 RJMCMC Algorithm Outline

Description

In order to outline the RJMCMC algorithm, we have closely followed Arnold et al. (2010). MCMC methods are a class of algorithms for sampling from a probability distribution (target distribution) based on constructing a Markov chain that has the target distribution as its equilibrium distribution (see e.g. a review of these methods in Gilks et al. (1996)). In our row clustering approach, the number of components is the number of row groups R and thus the target probability distribution is the posterior distribution $p(\Omega \mid \boldsymbol{Y}, R)$ of the parameter vector Ω as a current state given data $\boldsymbol{Y}$ and number of groups R (see eq. (7.12)). In particular, the fixed-dimensional M-H algorithm is a MCMC method that simulates a new state of parameters Ω^* from a proposal distribution $q(\Omega^* \mid \Omega, \boldsymbol{Y}, R)$ (see Section 7.1.5) given the current state Ω, data $\boldsymbol{Y}$ and number of groups R. This new state Ω^* is accepted with probability $\alpha_{\text{MH}} = \min(1, r_{\text{MH}})$ where r_{MH} is the acceptance ratio defined as

$$r_{\text{MH}} = \frac{p(\Omega^* \mid \boldsymbol{Y}, R)}{p(\Omega \mid \boldsymbol{Y}, R)} \frac{q(\Omega \mid \Omega^*, \boldsymbol{Y}, R)}{q(\Omega^* \mid \Omega, \boldsymbol{Y}, R)} = \frac{p(\boldsymbol{Y} \mid \Omega^*, R)}{p(\boldsymbol{Y} \mid \Omega, R)} \frac{p(\Omega^* \mid R)}{p(\Omega \mid R)} \frac{q(\Omega \mid \Omega^*, \boldsymbol{Y}, R)}{q(\Omega^* \mid \Omega, \boldsymbol{Y}, R)}.$$

Note that this ratio is the product of a likelihood ratio ($\text{LR}_{\text{MH}} = p(\boldsymbol{Y} \mid \Omega^*, R)/p(\boldsymbol{Y} \mid \Omega, R)$), a prior distribution ratio ($\text{PR}_{\text{MH}} = p(\Omega^* \mid R)/p(\Omega \mid R)$) and a proposal distribution ratio ($\text{QR}_{\text{MH}} = q(\Omega \mid \Omega^*, \boldsymbol{Y}, R)/q(\Omega^* \mid \Omega, \boldsymbol{Y}, R)$). Another feature is the *reversibility* of QR_{MH} which involves not only the probability of the proposed move $q(\Omega^* \mid \Omega, \boldsymbol{Y}, R)$ but also that of its reverse move $q(\Omega \mid \Omega^*, \boldsymbol{Y}, R)$. We accept the new state Ω^* from the Markov chain if a random draw from an $\mathcal{U}(0, 1)$ distribution is less than probability α_{MH}. Otherwise, the new state is rejected and the current state Ω becomes the new state for the chain. Details of the Metropolis-Hasting method for our approach are described in detail in Section 7.2.

The RJMCMC algorithm was introduced by Green (1995). This method is an extension to MCMC methodology which treats the number of groups R as an unknown parameter of interest and is able to propose the increase or reduction

of a single group. Therefore, this technique is a random sweep M-H algorithm adapted for examining the space of models of different dimensions. Green (1995) gave a comprehensive description of this algorithm. However, an outline of the RJMCMC algorithm for our one-dimensional clustering approach is as follows. We suppose the current state Ω (which now includes the number of clusters R in the parameter vector Ω) corresponds to a model $\mathcal{M}_\Omega$ which has dimension $\mathcal{D}_\Omega$, and the proposed state Ω^* is the parameter vector for a model $\mathcal{M}_{\Omega^*}$ with dimension $\mathcal{D}_{\Omega^*}$. Note that $\mathcal{M}_\Omega$ and $\mathcal{M}_{\Omega^*}$ can have identical dimensions but they might also differ when the proposed state involves an increase or a reduction in the number of clusters R. Thus, in order to propose a new state Ω^* the generation of $\ell \geq 0$ additional random variates u are required, where $\mathcal{D}_\Omega + \ell \geq \mathcal{D}_{\Omega^*}$. In the same manner, the reverse move (from Ω^* to Ω) also necessitates the generation of $\ell' \geq 0$ random variates u'. These variates u and u' are used to match dimensions. Once the random variates u and u' are drawn, the formulation of functions $B_{\Omega\to\Omega^*}(\Omega, u \mid \mathcal{M}_\Omega, \mathcal{M}_{\Omega^*})$ and its inverse $B_{\Omega^*\to\Omega}(\Omega^*, u' \mid \mathcal{M}_{\Omega^*}, \mathcal{M}_\Omega)$ is developed such that the mapping between (Ω, u) and (Ω^*, u') defined by $(\Omega^*, u') = B_{\Omega\to\Omega^*}(\Omega, u \mid \mathcal{M}_\Omega, \mathcal{M}_{\Omega^*})$ and $(\Omega, u) = B_{\Omega^*\to\Omega}(\Omega^*, u' \mid \mathcal{M}_{\Omega^*}, \mathcal{M}_\Omega)$ is diffeomorphism, i.e. differentiable bijection map whose inverse is also differentiable.

In order to achieve the detailed balance required in order for the Markov chain to correctly generate samples from the posterior distribution (target distribution), the RJMCMC sampler has to be *reversible*. Using an updated Metropolis-Hastings move allows us to accomplish this requirement when the move does not imply changes in dimensionality. In the case of a move where the dimensionality of the model varies, we must match dimensions $\mathcal{D}_\Omega + \ell = \mathcal{D}_{\Omega^*} + \ell'$ and the acceptance ratio r_{MH} is modified as

$$
\begin{aligned}
r_{\mathrm{RJ}} &= \frac{p(\Omega^*, \mathcal{M}_{\Omega^*} \mid \boldsymbol{Y})}{p(\Omega, \mathcal{M}_\Omega \mid \boldsymbol{Y})} \frac{q(u', \mathcal{M}_\Omega \mid \Omega^*, \mathcal{M}_{\Omega^*}, \boldsymbol{Y})}{q(u, \mathcal{M}_{\Omega^*} \mid \Omega, \mathcal{M}_\Omega, \boldsymbol{Y})} \left| \frac{\partial(\Omega^*, u')}{\partial(\Omega, u)} \right| \\
&= \frac{p(\boldsymbol{Y} \mid \Omega^*, \mathcal{M}_{\Omega^*})}{p(\boldsymbol{Y} \mid \Omega, \mathcal{M}_\Omega)} \frac{p(\Omega^* \mid \mathcal{M}_{\Omega^*})}{p(\Omega \mid \mathcal{M}_\Omega)} \frac{p(\mathcal{M}_{\Omega^*})}{p(\mathcal{M}_\Omega)} \frac{q(u' \mid \mathcal{M}_\Omega, \Omega^*, \mathcal{M}_{\Omega^*}, \boldsymbol{Y})}{q(u \mid \mathcal{M}_{\Omega^*}, \Omega, \mathcal{M}_\Omega, \boldsymbol{Y})} \\
&\quad \times \frac{g(\mathcal{M}_\Omega \mid \Omega^*, \mathcal{M}_{\Omega^*}, \boldsymbol{Y})}{g(\mathcal{M}_{\Omega^*} \mid \Omega, \mathcal{M}_\Omega, \boldsymbol{Y})} \left| \frac{\partial(\Omega^*, u')}{\partial(\Omega, u)} \right|
\end{aligned}
$$

where $g(\mathcal{M}_{\Omega^*} \mid \Omega, \mathcal{M}_\Omega, \boldsymbol{Y})$ is the probability of moving to model $\mathcal{M}_{\Omega^*}$ given the data $\boldsymbol{Y}$ and the current model $\mathcal{M}_\Omega$ with parameter vector Ω and $|\cdot|$ is the Jacobian matrix determinant from the change of dimensionality. Note that r_{RJ} can be seen

as the multiplication of the rates

$$r_{\mathrm{RJ}} = \mathrm{LR}_{\mathrm{RJ}} \times \mathrm{PR}_{\mathrm{RJ}} \times \mathrm{MPR}_{\mathrm{RJ}} \times \mathrm{QR}_{\mathrm{RJ}} \times \mathrm{NMPR}_{\mathrm{RJ}} \times \mathrm{J}_{\mathrm{RJ}} \tag{8.1}$$

where $\mathrm{LR}_{\mathrm{RJ}} = p(\boldsymbol{Y} \mid \Omega^*, \mathcal{M}_{\Omega^*})/p(\boldsymbol{Y} \mid \Omega, \mathcal{M}_{\Omega})$ is the likelihood ratio, $\mathrm{PR}_{\mathrm{RJ}} = p(\Omega^* \mid \mathcal{M}_{\Omega^*})/p(\Omega \mid \mathcal{M}_{\Omega})$ is the parameter prior distribution ratio, $\mathrm{MPR}_{\mathrm{RJ}} = p(\mathcal{M}_{\Omega^*})/p(\mathcal{M}_{\Omega})$ is the model prior ratio,
$\mathrm{QR}_{\mathrm{RJ}} = q(u' \mid \mathcal{M}_{\Omega}, \Omega^*, \mathcal{M}_{\Omega^*}, \boldsymbol{Y})/q(u \mid \mathcal{M}_{\Omega^*}, \Omega, \mathcal{M}_{\Omega}, \boldsymbol{Y})$ is the proposal distribution ratio, the probability to move to a new model ratio is $\mathrm{NMPR}_{\mathrm{RJ}} = g(\mathcal{M}_{\Omega} \mid \Omega^*, \mathcal{M}_{\Omega^*}, \boldsymbol{Y})/g(\mathcal{M}_{\Omega^*} \mid \Omega, \mathcal{M}_{\Omega}, \boldsymbol{Y})$, and $\mathrm{J}_{\mathrm{RJ}} = \left| \frac{\partial(\Omega^*, u')}{\partial(\Omega, u)} \right|$ is the Jacobian matrix determinant mapping between (Ω, u) and (Ω^*, u').

RJMCMC sampler

The general RJMCMC sampler is outlined as follows:

1. Specify an arbitrary initial value for the state $\Omega^{(0)} \mid \mathcal{M}_{\Omega}$ for which $p(\Omega^{(0)} \mid \boldsymbol{Y}, \mathcal{M}_{\Omega}) > 0$ where $\boldsymbol{Y}$ is the data set and $\mathcal{M}_{\Omega}$ is the current model.

2. At t^{th} iteration ($t = 1, 2, \ldots, T$):

   (a) Select a model $\mathcal{M}_{\Omega^*}$ with probability $g(\mathcal{M}_{\Omega^*} \mid \Omega, \mathcal{M}_{\Omega}, \boldsymbol{Y})$ which is related to a proposed state Ω^* .

   (b) If $\mathcal{M}_{\Omega^*} = \mathcal{M}_{\Omega^{(t-1)}}$ (i.e. there is no change in model), then

      i. Update the parameters from the current state $\Omega^{(t-1)}$ by using a sweep from the Metropolis-Hastings algorithm (see Section 7.1.5). It provides a new state $\Omega^* = \Omega^{\mathrm{MH}}$.

      ii. Set $\Omega^{(t)} = \Omega^*$ and return to step 2.

   (c) Otherwise ($\mathcal{M}_{\Omega^*} \neq \mathcal{M}_{\Omega^{(t-1)}}$), then

      i. Generate the random variates $u^{(t-1)}$ from a continuous and discrete distributions as appropriate.

      ii. Set $(\Omega^*, u^*) = B_{\Omega^{(t-1)} \to \Omega^*}(\Omega^{(t-1)}, u^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}}, \mathcal{M}_{\Omega^*})$.

      iii. Calculate the acceptance probability $\alpha_{\mathrm{RJ}} = \min(1, r_{\mathrm{RJ}})$ where the acceptance ratio r_{RJ} is defined as eq. (8.1).

iv. Accept the model $\mathcal{M}_{\Omega^*}$ and set $\Omega^{(t)} = \Omega^*$ and $\mathcal{M}_{\Omega^{(t)}} = \mathcal{M}_{\Omega^*}$ with probability α_{RJ}. Otherwise, reject $\mathcal{M}_{\Omega^*}$ and Ω^* and set $\Omega^{(t)} = \Omega^{(t-1)}$ and $\mathcal{M}_{\Omega^{(t)}} = \mathcal{M}_{\Omega^{(t-1)}}$.

3. Return the set of T samples $\{(\Omega^{(t)}, \mathcal{M}_{\Omega^{(t)}}), t = 1, \ldots, T\}$ from the posterior distribution $p(\Omega \mid \boldsymbol{Y})$.

4. Test whether convergence has been achieved. If not, increase T and return to step 2.

5. Summarise the set of T samples in a convenient way, such as computing posterior means, medians, credible intervals of the population size T, and in the posterior probability $p(\mathcal{M}_\Omega \mid \boldsymbol{Y})$ of each model $\mathcal{M}_\Omega$.

In the following section, the RJMCMC sampler to our approach is developed.

8.3 Application of the RJMCMC Sampler to Our Approach

In this section, the application of the RJMCMC sampler in the case of a one-dimensional clustering finite mixture-density model is illustrated. In particular, its application in the case of the row clustering model with unknown row membership probability (incomplete data). This model was described in Section 2.3. The analysis for the column clustering version is basically the same, but just replacing parameters related to rows with the equivalent column parameters. For the case of biclustering, the development may be a future research direction to explore.

Bayesian Estimation for Row Clustering Model

To estimate the parameters for the row clustering model, we have adopted a fully Bayesian approach based on M-H and RJMCMC samplers. We put priors on all unknown parameters and hyperpriors when necessary. Figure 8.1 depicts the directed acyclic graph (DAG) describing this fully Bayesian approach estimation for the row clustering model. The graph shows the priors, hyperparameters and their relationships, which are detailed in the following subsection.

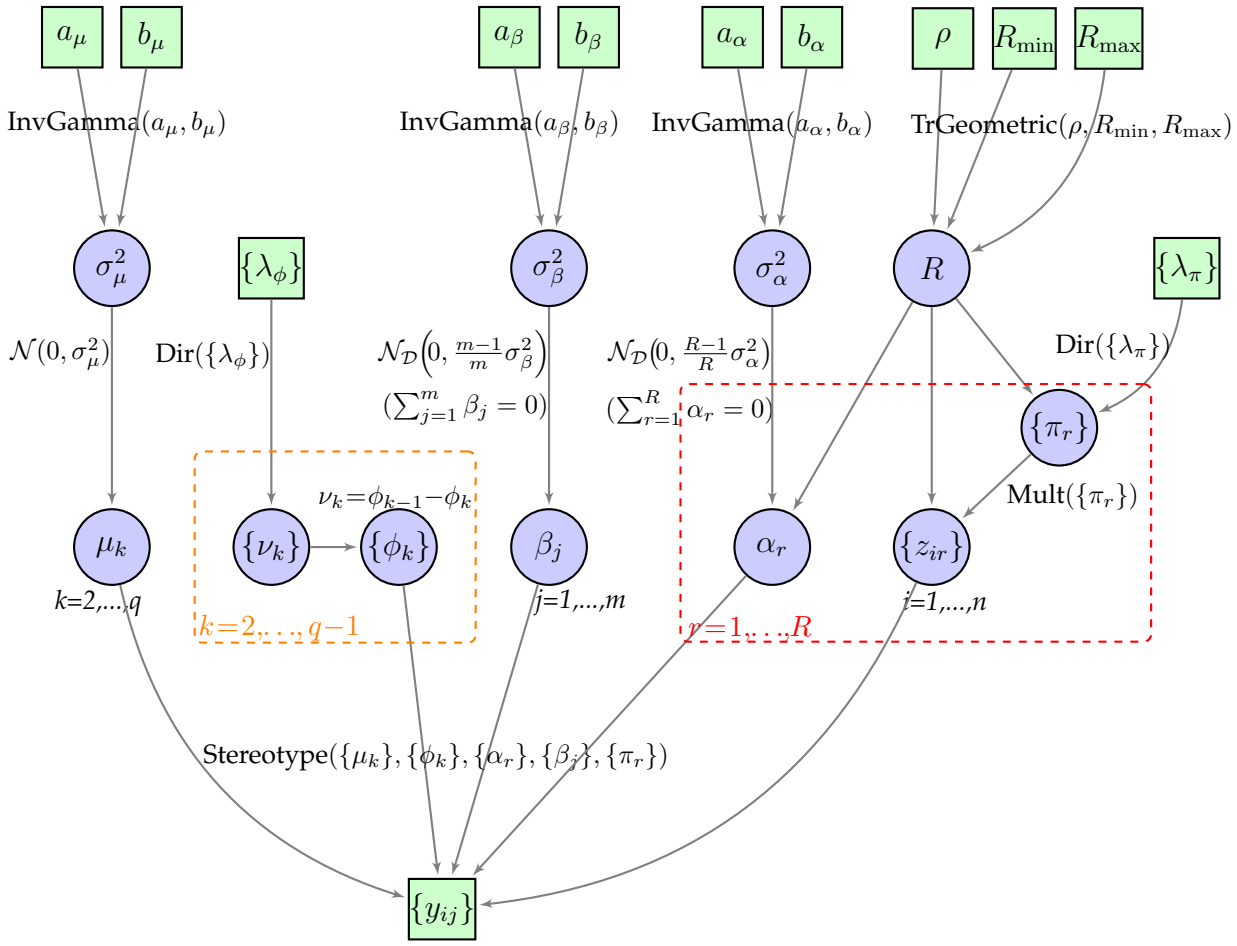

Figure 8.1: ***Directed acyclic graph:*** *Hierarchical Stereotype Mixture Model. Row Clustering. "TrGeometric" refers to a truncated Geometric distribution.*

Each step in the RJMCMC simulation is the result of one of these three moves: "update", "birth" and "death". In our case, we use a "split" move as a particular case of the birth move and a "merge" move as a specific death move in the corresponding parameters. We fully explain these two moves below. The parameters to estimate can be grouped into two main sets based on their dimensionality:

1. *Cut point ($\{\mu_k\}$), score ($\{\phi_k\}$), and column effect ($\{\beta_j\}$) parameters*: They have a fixed dimension (the number of categories q for $\{\mu_k\}$ and $\{\phi_k\}$ and the number of columns m for $\{\beta_j\}$) which do not vary during the sampling. In order to update these parameters, we use a sweep of the M-H algorithm. This does not alter the dimension of the parameter vector.

2. *Row cluster effect ($\{\alpha_r\}$) and row membership probability ($\{\pi_r\}$) parameters*: They

may have a variable dimension as the number of row clusters R may be altered during the RJMCMC sampling. In this case, the use of a reversible jump step is necessary and therefore the split and merge moves apply to these parameters. The value of R changes by 1 and the dimension of the parameter vector is altered. On the other hand, if the move does not alter the number of clusters R (therefore the move is within the same family class), then we apply an update move by using a sweep from the M-H algorithm.

Update Move. Prior Distributions and Hyperparameters

The update move for each parameter from the row clustering model is defined as a simple M-H move. We adopt the same prior distributions and their related hyperparameters as they were described in Section 7.2.2. The prior distributions for the model are listed in Table 8.1 with the default values of the relevant defining hyperparameters given alongside. These default setting can be altered to suit another particular problem. All priors are proper.

Table 8.1: RJMCMC sampler. Priors and default settings for the hyperparameters defining their distributions.

Parameter	Prior Distribution	Hyperparameters	Notes
σ_μ^2	InverseGamma $\left(\nu_{\sigma_\mu}, \delta_{\sigma_\mu}\right)$	$\nu_{\sigma_\mu} = 3$ $\delta_{\sigma_\mu} = 40$	
$\{\mu_k\}$	$\mathcal{N}(0, \sigma_\mu^2)$		
$\{\phi_k\}$	$\{\nu_k\} \sim$ Dirichlet($\boldsymbol{\lambda}_\phi$) $\nu_k = \phi_{k+1} - \phi_k$	$\boldsymbol{\lambda}_\phi = \mathbf{1}$	
σ_α^2	InverseGamma $\left(\nu_{\sigma_\alpha}, \delta_{\sigma_\alpha}\right)$	$\nu_{\sigma_\alpha} = 3$ $\delta_{\sigma_\alpha} = 40$	
$\{\alpha_r\}$	DegenNormal$(R; 0, \sigma_\alpha^2)$		$\sum_{r=1}^{R} \alpha_r = 0$
σ_β^2	InverseGamma $\left(\nu_{\sigma_\beta}, \delta_{\sigma_\beta}\right)$	$\nu_{\sigma_\beta} = 3$ $\delta_{\sigma_\beta} = 40$	
$\{\beta_j\}$	DegenNormal$(m; 0, \sigma_\beta^2)$		$\sum_{j=1}^{m} \beta_j = 0$
$\{\gamma_{rj}\}$	DegenNormal$(R, m; 0, \sigma_\gamma^2)$	$\sigma_\gamma^2 = 5$	$\sum_{r=1}^{R} \gamma_{rj} = \sum_{j=1}^{m} \gamma_{rj} = 0$
$\{\pi_r\}$	Dirichlet($\boldsymbol{\lambda}_\pi$)	$\boldsymbol{\lambda}_\pi = \mathbf{1}$	$\sum_{r=1}^{R} \pi_r = 1$
R	TrGeometric $(1-\rho, R_{\min}, R_{\max})$	$\rho = 0.8$ $R_{\min} = 1,\ R_{\max} = 10$	$R_{\min} \leq R \leq R_{\max}$

RJMCMC Sweep

In implementing the RJMCMC sampler in this model we break the parameters into three blocks. Each sweep of the sampler updates each block in sequence. For blocks 1 and 2, the sampler sequentially updates one parameter at a time leaving the other components unchanged. For block 3 a choice is made at random among update, split and merge moves. The first block contains the hyperparameters for the prior distributions, which can be updated by using Metropolis-Hastings proposals with acceptance ratio equal to 1 (Gibbs sampling). The second block relates to the cut point $\{\mu_k\}$, score $\{\phi_k\}$ and column effect $\{\beta_j\}$ parameters. They can be updated using a M-H sweep as they have a fixed dimension and they do not increase or decrease when the sampler moves from one model to another model with different numbers of clusters R. In the third block, the focus is on the row cluster $\{\alpha_r\}$ and row membership $\{\pi_r\}$ parameters and there are three different move options:

- an *update* move when it does not alter the number of clusters R,
- *split* a randomly chosen cluster into two, and
- *merge* a randomly selected neighbouring pair of clusters into a single cluster.

As noted above, an *update* move is a conventional MCMC move, which is made by Metropolis-Hastings proposals. *Split* and *merge* moves form a reversible pair of dimension changing moves which take place using a reversible jump step. See more details about these two moves below. Summary details of the different move types and default parameter settings are collected together in Table 8.2. The sampler always run a M-H step when a parameter from block 1 or 2 is chosen randomly (indicated as Pr(Move)=1 in Table 8.2). However, the moves in the parameters from the third block have assigned probabilities as is explained in the following Section.

Probabilities of Dimension-Changing Moves

The moves in the third block (see Table 8.2) have assigned probabilities $(p_\alpha, p_\pi, p_{\text{split}}, p_{\text{merge}})$ where p_α and p_π are related to the *update* move for $\{\alpha_r\}$ and $\{\pi_r\}$ respectively. The probabilities $(p_{\text{split}}, p_{\text{merge}})$ are related to the *split* and *merge*

Table 8.2: Transition groups, move types, move probabilities and default settings for the hyperparameters controlling the proposal distributions for the RJMCMC sampler in the estimation of the row clustering model. The move types are labelled M-H=Metropolis-Hastings and RJ=Reversible jump. The split and merge moves are emphasized in bold.

Block	Move Param.	Prop. Constants	Pr(Move)	Move Type
1 Hyperpar.	σ^2_μ	$\nu_{\sigma_\mu} = 3\ \delta_{\sigma_\mu} = 40$	1	M-H
	σ^2_α	$\nu_{\sigma_\alpha} = 3\ \delta_{\sigma_\alpha} = 40$	1	M-H
	σ^2_β	$\nu_{\sigma_\beta} = 3\ \delta_{\sigma_\beta} = 40$	1	M-H
2 General Parameters	$\{\mu_k\}$	$\sigma^2_{\mu_p} = 0.3$	1	M-H
	$\{\phi_k\}$		1	M-H
	$\{\beta_j\}$	$\sigma^2_{\beta_p} = 0.3$	1	M-H
3 Row Parameters	$\{\alpha_r\}$	$\sigma^2_{\alpha_p} = 0.3$	$p_\alpha = \frac{1}{2}(1-p)$	M-H
	$\{\pi_r\}$	$\sigma^2_{\pi_p} = 0.3$	$p_\pi = \frac{1}{2}(1-p)$	M-H
	Split	$p = 0.3$	$p_{\text{split}} = p\frac{\rho}{1+\rho}$	**RJ**
	Merge	$\rho = 0.8$	$p_{\text{merge}} = p\frac{1}{1+\rho}$	**RJ**

moves respectively and are the probabilities of dimension-changing moves which we have formulated following Green (1995) as

$$\begin{cases} p_{\text{split}} = p\frac{\rho}{1+\rho}, & p_{\text{merge}} = p\frac{1}{1+\rho} & \text{for } R_{\min} < R < R_{\max} \\ p_{\text{split}} = p, & p_{\text{merge}} = 0 & \text{for } R = R_{\min} \\ p_{\text{split}} = 0, & p_{\text{merge}} = p & \text{for } R = R_{\max}, \end{cases}$$

where ρ is the parameter of the truncated Geometric prior for R (see Table 8.1) and p is the probability related to the proportion of times a split or merge move is proposed. This choice means that the probability of a dimension-changing move, i.e. $p_{\text{split}} + p_{\text{merge}}$, is p for $R_{\min} \leq R \leq R_{\max}$, and also implies that $\frac{p_{\text{merge}}}{p_{\text{split}}}\rho = 1$ for $R_{\min} < R < R_{\max}$. This latter property means that the proposal and prior ratios for R conveniently cancel in the construction of the acceptance ratio. The truncated Geometric prior distribution for the number of clusters R is cancelled out by the acceptance ratio for the dimension-changing moves (see development of $r^{\text{split}}_{\text{RJ}}$ and $r^{\text{merge}}_{\text{RJ}}$ below) depending only on the parameter ρ. Finally, we take the constant $p = 0.3$, which is large enough so that dimension-changing moves are proposed frequently, but not so large that the parameters within models of a given dimension do not have time to mix. The remaining probability of 0.7 is divided equally between the $\{\alpha_r\}$ and $\{\pi_r\}$ *update* moves.

Split and Merge Moves. Description

The split and merge moves consider an alteration of the dimensionality in the parameter space (Richardson and Green, 1997). These are particular birth and death moves. The merge move reduces the dimension of the model and the split move increases it. Choosing randomly between these moves proposes a new model. In the row clustering model, these characteristic RJ moves vary the row effect $\{\alpha_r\}$ and row membership $\{\pi_r\}$ parameters. They also alter the number of row clusters R which is considered a parameter in this scheme. The procedure to apply these moves must preserve both the *reversible* requirement and the constraints for identifiability reasons ($\sum_{r=1}^{R} \alpha_r = 0$ and $\sum_{r=1}^{R} \pi_r = 1$). At iteration t ($t \in \{1, 2, \ldots, T\}$), an outline of how we use those dimension-changing moves is:

- *Split move* :

  1. Draw randomly two uniform variates $\Delta_1, \Delta_2 \sim U(0, 1)$.
  2. Select randomly one cluster $r \in \{1, \ldots, R^{(t-1)}\}$.
  3. Generate two new parameters for α_r and π_r halving them from iteration $(t-1)$ as follows:

$$\begin{aligned} \alpha_r^{(t)} &= \Delta_1 \alpha_r^{(t-1)} \text{ and } \alpha_{r+1}^{(t)} = (1-\Delta_1)\alpha_r^{(t-1)}, \\ \pi_r^{(t)} &= \Delta_2 \pi_r^{(t-1)} \text{ and } \pi_{r+1}^{(t)} = (1-\Delta_2)\pi_r^{(t-1)}. \end{aligned} \tag{8.2}$$

  4. Increase the number of row clusters $R^{(t-1)}$ by 1: $R^{(t)} = R^{(t-1)} + 1$.
  5. Relabel $\{r+1, \ldots, R^{(t-1)}\}$ as $\{r+2, \ldots, R^{(t)}\}$ in $\{\alpha_r^{(t)}\}$ and $\{\pi_r^{(t)}\}$.

- *Merge move:*

  1. Choose randomly one cluster $r \in \{1, \ldots, R^{(t-1)} - 1\}$.
  2. Select the adjacent component $r+1$.
  3. Generate two new parameters merging the two components from iteration $(t-1)$ as:

$$\begin{aligned} \alpha_r^{(t)} &= \alpha_r^{(t-1)} + \alpha_{r+1}^{(t-1)}, \\ \pi_r^{(t)} &= \pi_r^{(t-1)} + \pi_{r+1}^{(t-1)}. \end{aligned} \tag{8.3}$$

  4. Reduce the number of row clusters $R^{(t-1)}$ by 1: $R^{(t)} = R^{(t-1)} - 1$.

5. Relabel $\{r+2,\ldots,R^{(t-1)}\}$ as $\{r+1,\ldots,R^{(t)}\}$ in $\{\alpha_r^{(t)}\}$ and $\{\pi_r^{(t)}\}$.

Split Move. Acceptance Ratio

A split move is accepted with the probability of $\alpha_{\text{split}} = \min(1, r_{\text{RJ}}^{\text{split}})$ where the acceptance ratio $r_{\text{RJ}}^{\text{split}}$ is defined as eq. (8.1) where

- Likelihood function ratio:

$$\text{LR}_{\text{RJ}}^{\text{split}} = \text{LR}_{\text{RJ}} = \frac{p(\boldsymbol{Y} \mid \Omega^{(t)}, \mathcal{M}_{\Omega^{(t)}})}{p(\boldsymbol{Y} \mid \Omega^{(t-1)}, \mathcal{M}_{\Omega^{(t-1)}})},$$

  where $\mathcal{M}_{\Omega^{(t)}}$ is the model after splitting and $\mathcal{M}_{\Omega^{(t-1)}}$ is the model before splitting.

- Prior distribution ratio:

$$\begin{aligned}\text{PR}_{\text{RJ}}^{\text{split}} &= \text{PR}_{\alpha}^{\text{split}} \times \text{PR}_{\pi}^{\text{split}} \times \text{PR}_{R}^{\text{split}} \\ &= \frac{p(\alpha^{(t)} \mid \mathcal{M}_{\Omega^{(t)}})}{p(\alpha^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}})} \frac{p(\pi^{(t)} \mid \mathcal{M}_{\Omega^{(t)}})}{p(\pi^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}})} \frac{p(R^{(t)} \mid \mathcal{M}_{\Omega^{(t)}})}{p(R^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}})},\end{aligned}$$

  where

$$\begin{aligned}p(\{\alpha^{(t)}\} \mid \mathcal{M}_{\Omega^{(t)}}) &= p(\alpha_r^{(t)}, \alpha_{r+1}^{(t)} \mid \mathcal{M}_{\Omega^{(t)}}) \\ &= \frac{1}{\sqrt{2\pi}\sigma_\alpha^2} \exp\left\{-\frac{1}{2\sigma_\alpha^2}(\alpha_r^{(t)^2} + \alpha_{r+1}^{(t)^2})\right\}, \\ p(\{\alpha^{(t-1)}\} \mid \mathcal{M}_{\Omega^{(t-1)}}) &= p(\alpha_r^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}}) = \frac{1}{\sqrt{2\pi}\sigma_\alpha^2} \exp\left\{-\frac{1}{2\sigma_\alpha^2}\alpha_r^{(t-1)^2}\right\}, \\ p(\{\pi^{(t)}\} \mid \mathcal{M}_{\Omega^{(t)}}) &= \frac{\Gamma(R^{(t)}\lambda_\pi)}{R^{(t)}\Gamma(\lambda_\pi)} \prod_{r=1}^{R^{(t)}} (\pi_r^{(t)})^{\lambda_\pi - 1} \qquad \text{with } R^{(t)} = R^{(t-1)} + 1, \\ p(\{\pi^{(t-1)}\} \mid \mathcal{M}_{\Omega^{(t-1)}}) &= \frac{\Gamma(R^{(t-1)}\lambda_\pi)}{R^{(t-1)}\Gamma(\lambda_\pi)} \prod_{r=1}^{R^{(t-1)}} (\pi_r^{(t-1)})^{\lambda_\pi - 1}, \\ p(R^{(t)} \mid \mathcal{M}_{\Omega^{(t)}}) &= 1 \;\text{ and }\; p(R^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}}) = 1.\end{aligned}$$

Therefore,

$$\mathrm{PR}_{\alpha}^{\mathrm{split}} = \exp\left\{-\frac{1}{2\sigma_{\alpha}^{2}}\left(\alpha_{r}^{(t)^{2}} + \alpha_{r+1}^{(t)^{2}} - \alpha_{r}^{(t-1)^{2}}\right)\right\},$$

$$\mathrm{PR}_{\pi}^{\mathrm{split}} = \frac{\Gamma(R^{(t)}\lambda_{\pi})}{\Gamma(R^{(t-1)}\lambda_{\pi})\Gamma(\lambda_{\pi})}\left(\frac{\pi_{r}^{(t)}\pi_{r+1}^{(t)}}{\pi_{r}^{(t-1)}}\right)^{\lambda_{\pi}-1}, \text{and } \mathrm{PR}_{R}^{\mathrm{split}} = 1.$$

- Model prior ratio:

$$\mathrm{MPR}_{\mathrm{RJ}}^{\mathrm{split}} = \frac{p(\mathcal{M}_{\Omega^{(t)}})}{p(\mathcal{M}_{\Omega^{(t-1)}})} = \frac{p(R^{(t)})}{p(R^{(t-1)})} = \frac{\frac{(1-\rho)\rho^{R^{(t)}-1}}{\rho^{R_{\min}}-\rho^{R_{\max}}}}{\frac{(1-\rho)\rho^{R^{(t-1)}-1}}{\rho^{R_{\min}}-\rho^{R_{\max}}}} = \frac{\rho^{R^{(t)}-1}}{\rho^{R^{(t-1)}-1}} = \rho.$$

- Proposal distribution ratio:

  We define $u^{(t-1)} = \{\Delta_1, \Delta_2\}$ and $u^{(t)} = \varnothing$, where $\Delta_1, \Delta_2 \sim U(0,1)$. Therefore,

$$\mathrm{QR}_{\mathrm{RJ}}^{\mathrm{split}} = \frac{q(u^{(t)} \mid \mathcal{M}_{\Omega^{(t-1)}}, \Omega^{(t)}, \mathcal{M}_{\Omega^{(t)}}, \boldsymbol{Y})}{q(u^{(t-1)} \mid \mathcal{M}_{\Omega^{(t)}}, \Omega, \mathcal{M}_{\Omega^{(t-1)}}, \boldsymbol{Y})} = 1.$$

- Probability to move to a new model ratio:

$$\mathrm{NMPR}_{\mathrm{RJ}}^{\mathrm{split}} = \frac{g(\mathcal{M}_{\Omega^{(t-1)}} \mid \Omega^{(t)}, \mathcal{M}_{\Omega^{(t)}}, \boldsymbol{Y})}{g(\mathcal{M}_{\Omega^{(t)}} \mid \Omega^{(t-1)}, \mathcal{M}_{\Omega^{(t-1)}}, \boldsymbol{Y})} = \frac{p_{\mathrm{merge}}}{p_{\mathrm{split}}} = \frac{1}{\rho},$$

  for $R_{\min} \leq R \leq R_{\max}$.

- Jacobian determinant between $(\alpha_r^{(t-1)}, \Delta_1, \pi_r^{(t-1)}, \Delta_2)$ and $(\alpha_r^{(t)}, \pi_r^{(t)}, \alpha_{r+1}^{(t)}, \pi_{r+1}^{(t)})$ where $\Delta_1, \Delta_2 \sim U(0,1)$ and $(\alpha_r^{(t)}, \alpha_{r+1}^{(t)}, \pi_r^{(t)}, \pi_{r+1}^{(t)})$ are defined as equation (8.2):

$$\mathrm{J}_{\mathrm{RJ}}^{\mathrm{split}} = \left|\frac{\partial\left(\alpha_r^{(t-1)}, \Delta_1, \pi_r^{(t-1)}, \Delta_2\right)}{\partial\left(\alpha_r^{(t)}, \pi_r^{(t)}, \alpha_{r+1}^{(t)}, \pi_{r+1}^{(t)}\right)}\right| = \begin{vmatrix} \Delta_1 & 1-\Delta_1 & 0 & 0 \\ \alpha_r^{(t-1)} & -\alpha_r^{(t-1)} & 0 & 0 \\ 0 & 0 & \Delta_2 & 1-\Delta_2 \\ 0 & 0 & \pi_r^{(t-1)} & -\pi_r^{(t-1)} \end{vmatrix}$$

$$= \alpha_r^{(t-1)}\pi_r^{(t-1)}.$$

Merge Move. Acceptance Ratio

A merge move is accepted with the probability of $\alpha_{\text{merge}} = \min(1, r_{\text{RJ}}^{\text{merge}}) = \min(1, r_{\text{RJ}}^{\text{split}^{-1}})$ where the acceptance ratio $r_{\text{RJ}}^{\text{merge}}$ is defined as eq. (8.1) where

- Likelihood function ratio:

$$\text{LR}_{\text{RJ}}^{\text{merge}} = \text{LR}_{\text{RJ}} = \frac{p(\boldsymbol{Y} \mid \Omega^{(t)}, \mathcal{M}_{\Omega^{(t)}})}{p(\boldsymbol{Y} \mid \Omega^{(t-1)}, \mathcal{M}_{\Omega^{(t-1)}})},$$

  where $\mathcal{M}_{\Omega^{(t)}}$ is the model after merging and $\mathcal{M}_{\Omega^{(t-1)}}$ is the model before merging.

- Prior distribution ratio:

$$\begin{aligned}\text{PR}_{\text{RJ}}^{\text{merge}} &= \text{PR}_{\alpha}^{\text{merge}} \times \text{PR}_{\pi}^{\text{merge}} \times \text{PR}_{R}^{\text{merge}} \\ &= \frac{p(\alpha^{(t)} \mid \mathcal{M}_{\Omega^{(t)}})}{p(\alpha^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}})} \frac{p(\pi^{(t)} \mid \mathcal{M}_{\Omega^{(t)}})}{p(\pi^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}})} \frac{p(R^{(t)} \mid \mathcal{M}_{\Omega^{(t)}})}{p(R^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}})},\end{aligned}$$

  where

$$\begin{aligned}p(\{\alpha^{(t)}\} \mid \mathcal{M}_{\Omega^{(t)}}) &= p(\alpha_r^{(t)} \mid \mathcal{M}_{\Omega^{(t)}}) = \frac{1}{\sqrt{2\pi}\sigma_\alpha^2} \exp\left\{-\frac{1}{2\sigma_\alpha^2}\alpha_r^{(t)^2}\right\}, \\ p(\{\alpha^{(t-1)}\} \mid \mathcal{M}_{\Omega^{(t-1)}}) &= p(\alpha_r^{(t-1)}, \alpha_{r+1}^{(t-1)} \mid \mathcal{M}_{\Omega^{(t)}}) \\ &= \frac{1}{\sqrt{2\pi}\sigma_\alpha^2} \exp\left\{-\frac{1}{2\sigma_\alpha^2}(\alpha_r^{(t-1)^2} + \alpha_{r+1}^{(t-1)^2})\right\}, \\ p(\{\pi^{(t)}\} \mid \mathcal{M}_{\Omega^{(t)}}) &= \frac{\Gamma(R^{(t)}\lambda_\pi)}{R^{(t)}\Gamma(\lambda_\pi)} \prod_{r=1}^{R^{(t)}} (\pi_r^{(t)})^{\lambda_\pi - 1} \qquad \text{with } R^{(t)} = R^{(t-1)} - 1, \\ p(\{\pi^{(t-1)}\} \mid \mathcal{M}_{\Omega^{(t-1)}}) &= \frac{\Gamma(R^{(t-1)}\lambda_\pi)}{R^{(t-1)}\Gamma(\lambda_\pi)} \prod_{r=1}^{R^{(t-1)}} (\pi_r^{(t-1)})^{\lambda_\pi - 1}, \\ p(R^{(t)} \mid \mathcal{M}_{\Omega^{(t)}}) &= 1 \;\text{ and }\; p(R^{(t-1)} \mid \mathcal{M}_{\Omega^{(t-1)}}) = 1.\end{aligned}$$

  Therefore,

$$\begin{aligned}\text{PR}_{\alpha}^{\text{merge}} &= \exp\left\{-\frac{1}{2\sigma_\alpha^2}\left(\alpha_r^{(t)^2} - \alpha_r^{(t-1)^2} - \alpha_{r+1}^{(t-1)^2}\right)\right\}, \\ \text{PR}_{\pi}^{\text{merge}} &= \frac{\Gamma(R^{(t)}\lambda_\pi)\Gamma(\lambda_\pi)}{\Gamma(R^{(t-1)}\lambda_\pi)} \left(\frac{\pi_r^{(t)}}{\pi_r^{(t-1)}\pi_{r+1}^{(t-1)}}\right)^{\lambda_\pi - 1}, \text{and } \text{PR}_{R}^{\text{merge}} = 1.\end{aligned}$$

- Model prior ratio:

$$\text{MPR}_{\text{RJ}}^{\text{merge}} = \frac{p(\mathcal{M}_{\Omega^{(t)}})}{p(\mathcal{M}_{\Omega^{(t-1)}})} = \frac{p(R^{(t)})}{p(R^{(t-1)})} = \frac{\frac{(1-\rho)\rho^{R^{(t)}-1}}{\rho^{R_{\min}}-\rho^{R_{\max}}}}{\frac{(1-\rho)\rho^{R^{(t-1)}-1}}{\rho^{R_{\min}}-\rho^{R_{\max}}}} = \frac{\rho^{R^{(t-1)}-2}}{\rho^{R^{(t-1)}-1}} = \frac{1}{\rho}.$$

- Proposal distribution ratio:

  We define $u^{(t-1)} = \varnothing$ and $u^{(t)} = \{\Delta_1, \Delta_2\}$ where $\Delta_1, \Delta_2 \sim U(0,1)$. Therefore,

$$\text{QR}_{\text{RJ}}^{\text{merge}} = \frac{q(u^{(t)} \mid \mathcal{M}_{\Omega^{(t-1)}}, \Omega^{(t)}, \mathcal{M}_{\Omega^{(t)}}, \boldsymbol{Y})}{q(u^{(t-1)} \mid \mathcal{M}_{\Omega^{(t)}}, \Omega, \mathcal{M}_{\Omega^{(t-1)}}, \boldsymbol{Y})} = 1.$$

- Probability to move to a new model ratio:

$$\text{NMPR}_{\text{RJ}}^{\text{merge}} = \frac{g(\mathcal{M}_{\Omega^{(t-1)}} \mid \Omega^{(t)}, \mathcal{M}_{\Omega^{(t)}}, \boldsymbol{Y})}{g(\mathcal{M}_{\Omega^{(t)}} \mid \Omega^{(t-1)}, \mathcal{M}_{\Omega^{(t-1)}}, \boldsymbol{Y})} = \frac{p_{\text{split}}}{p_{\text{merge}}} = \rho,$$

  for $R_{\min} \leq R \leq R_{\max}$.

- Jacobian determinant between $(\alpha_r^{(t-1)}, \pi_r^{(t-1)}, \alpha_{r+1}^{(t-1)}, \pi_{r+1}^{(t-1)})$ and $(\alpha_r^{(t)}, \Delta_1, \pi_r^{(t)}, \Delta_2)$ where $\Delta_1 = \frac{\alpha_r^{(t-1)}}{\alpha_r^{(t)}}$, $\Delta_2 = \frac{\pi_r^{(t-1)}}{\pi_r^{(t)}}$ and $(\alpha_r^{(t)}, \alpha_{r+1}^{(t)}, \pi_r^{(t)}, \pi_{r+1}^{(t)})$ are defined as equation (8.3):

$$\text{J}_{\text{RJ}}^{\text{merge}} = \left| \frac{\partial \left(\alpha_r^{(t)}, \Delta_1^{(t)}, \pi_r^{(t)}, \Delta_2^{(t)} \right)}{\partial \left(\alpha_r^{(t-1)}, \pi_r^{(t-1)}, \alpha_{r+1}^{(t-1)}, \pi_{r+1}^{(t-1)} \right)} \right| = \begin{vmatrix} 1 & 1 & 0 & 0 \\ \frac{1}{\alpha_r^{(t)}} & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & \frac{1}{\pi_r^{(t)}} & 0 \end{vmatrix} = \frac{1}{\alpha_r^{(t)} \pi_r^{(t)}}.$$

There are two final considerations for our RJMCMC approach. Firstly, the application of a label switching procedure is required as we are dealing with finite mixtures (see Section 7.2.5). Secondly, the use of a specific technique to diagnose convergence for RJMCMC as we will describe in the following section.

8.4 Convergence Diagnostic for RJMCMC Samplers

Monitoring of MCMC convergence on the basis of empirical statistics of the chain is important, although not a substitute for a good theoretical understanding of

the chain. There are several well-known techniques to diagnose convergence for a fixed-dimensional MCMC sampler as discussed in Section 7.1.3. A major difficulty arises when diagnosing convergence applies to a MCMC sampler such as the RJMCMC sampler, which involves jumping between models of different dimensions and therefore both the length and interpretation of the state vector may change from iteration to iteration. Green and Hastie (2009) stated that the degree of confidence that convergence has been achieved provided by "passing" a fixed-dimensional diagnostic convergence test declines very rapidly as the dimension of the state space changes.

Not many convergence diagnostic methods for RJMCMC are available. In finite mixtures, one 2-step strategy consists of monitoring the convergence of parameter related to the number of components in the mixture individually and, later, testing the convergence within each model individually after that (Richardson and Green, 1997). The drawback of this approach is that some models will not be visited very often even in long run samplings and therefore diagnostic convergence within those models is almost impossible (Brooks, 1997). An approach extending the Gelman and Rubin method (Gelman and Rubin (1992) and Brooks (1998), see Appendix F.2) to the RJMCMC case is given in Brooks and Giudici (2000). It is based on a 2-way ANOVA analysis of the simulation output over multiple chain replications for those parameters that do not change their interpretation as the sampler moves from model to model. In this framework, the potential scale reduction factor (PSRF) can be computed and monitored. The drawback is that those parameters may not characterize the whole set of parameters and therefore mislead the convergence assessments. Castelloe and Zimmerman (2002) modified the method proposed by Brooks and Giudici (2000) to a weighted 2-way ANOVA analysis with the weights being specified in proportion to the frequency of model visits. This approach prevents the PSRF being dominated by a few visits to rare models. The approach utilizes multiple chain replications and detects the following:

1. variation that is not homogeneous between chains (like in the Gelman and Rubin method),

2. between-model variation that differs from one chain to another, and

3. significant differences in the frequencies of model visits from one chain to another.

Alternative approaches are a nonparametric method for assessing RJMCMC sampler performance based on distance measures (Brooks et al., 2003) and a distance-based diagnostic when the underlying model can be formulated in the marked point process framework (Sisson and Fan, 2007).

In this thesis, the method proposed by Castelloe and Zimmerman (2002) to assess the convergence of RJMCMC sampler was implemented. Technical details can be found in the Section 4 of their paper and an outline of the method is described in Appendix I.1.

8.5 Simulation Study

We set up a simulation study to test how reliably we were able to estimate the parameters (including the number of clusters) for our row clustering approach by using the RJMCMC algorithm. Generally speaking, we simulate datasets and then run multiple RJMCMC samplers to fit the more appropriate model.

The design of the RJMCMC sampler refers to an ordinal response variable with four categories ($q = 4$) and we varied the number of row clusters ($R = 2, \ldots, 6$) in order to test if the RJMCMC sampler returns the correct number of clusters. The sample size ($n = 1000$) and the number of columns ($m = 3$) is fixed. For each number of row clusters R, a single set of parameters values was chosen and $H = 100$ data sets (replicates) were generated. For each data set, we assessed the convergence of the RJMCMC sampler by running $S = 10$ chains in parallel from random starting points. We ran each chain for a initial 20000 iterations, but discarded these initial samples (burn-in period). We then ran each chain for a further 200000 updates, storing only every 10th state (thinning). We used the methods of Castelloe and Zimmerman (2002) to assess the convergence of these chains. For each chain, we summarised results computing the mean, median, interquartile range, standard deviation, time series standard error highest posterior density interval (HPD) and maximum a posteriori estimator (MAP) for the free parameter vector Ω. Additionally, we computed the proportion of times where the estimation agrees with true parameters over the total number of chains. The simulation study procedure for the RJMCMC with row clustering case is outlined in the following steps:

Step 1. Model specification

One row clustering without interaction model ($\mu_k + \phi_k(\alpha_r + \beta_j)$) is specified based on:

- Set $R_{\min} = 2$, $R_{\max} = 6$.
- Select the number of clusters R from the set $\{R_{\min}, \ldots, R_{\max}\}$ (1 option).
  R acts as the true number of row clusters.
  This fixes the row effects $\{\alpha_1, \ldots, \alpha_R\}$ (with $\sum_{r=1}^{R} \alpha_r = 0$) and
  the prior mixing probabilities $\{\pi_1, \ldots, \pi_R\}$ (with $\sum_{r=1}^{R} \pi_r = 1$).
- Select the number of response categories: $q = 4$ in all cases (1 option).
  This fixes $\{\mu_1, \ldots, \mu_q\}$ (with $\mu_1 = 0$) and
  the ordinal response cut levels $\phi_1 \leq \phi_2 \leq \ldots \leq \phi_q$ (with $\phi_1 = 0$ and $\phi_q = 1$).
- Select the number of columns: $m = 3$ in all cases (1 option).
  This fixes $\{\beta_1, \ldots, \beta_m\}$ with $\sum_{j=1}^{m} \beta_j = 0$).
- Set the sample size $n = 1000$ (1 option).

At the end of this step we know, for the chosen model:

- The number of row groups R.
- The number of response categories q.
- The number of columns m.
- The sample size n.
- The total number of free parameters $K = 2q + 2R + m - 6$.
- The parameter vector Ω consisting of free parameters:

$$\{\alpha_1, \ldots, \alpha_R\}, \{\beta_1, \ldots, \beta_m\}, \{\pi_1, \ldots, \pi_R\}, \{\mu_1, \ldots, \mu_q\}, \text{ and } \{\phi_1, \ldots, \phi_q\},$$

  and as a consequence we can calculate the values of the linear predictors

$$\eta_{krj} = \mu_k + \phi_k \left(\alpha_r + \beta_j\right),$$

  for $k \in \{1, \ldots, q\}$, $r \in \{1, \ldots, R\}$ and $j \in \{1, \ldots, m\}$.

Step 2. Simulator specification

Set the parameters for the simulator specifying:

- The number of replicates to run (i.e. distinct datasets): $H = 100$.
- The number of chains in each replicate: $S = 10$.

As a result we run 1000 RJMCMC chains.

Step 3. Markov Chain specification

Set the chain parameters specifying:

- The number of iterations in the burn-in period: nburn=20000.
- The number of iterations to store: nstore=20000.
- The thinning rate: nthin=10.

As a result we run each chain for a overall of T=nburn+(nthin $\times$ nstore)=220000 iterations.

Step 4. RJ move parameter and hyperparameter specification

Set the constant related to the proportion of times that a split or merge move is proposed: $p = 0.3$.
Set the hyperparameters values specifying:

- The parameters of a truncated Geometric distribution which is the prior for the number of row clusters R:
  - The range of number of possible row clusters: $R^{\text{RJ}}_{\text{min}} = 1$ and $R^{\text{RJ}}_{\text{max}} = 10$.
  - The success parameter: $1 - \rho = 0.2$.
- Shape and scale parameters to specify an Inverse Gamma distribution which is the prior for the standard deviation parameter from a Normal distribution related to
  - the cut point parameters $\{\mu_k\}$: $\nu_\mu = 3$, $\delta_\mu = 40$,
  - the row cluster parameters $\{\alpha_r\}$: $\nu_\alpha = 3$, $\delta_\alpha = 40$, and

  - the column parameters $\{\beta_j\}$: $\nu_\beta = 3$, $\delta_\beta = 40$.

- Parameter vector regarding a Dirichlet distribution related to
  - the score parameters $\{\phi_k\}$: $\boldsymbol{\lambda}_\phi = \mathbf{1}$, and
  - the prior mixing probabilities $\boldsymbol{\lambda}_\pi = \mathbf{1}$.

As a result a dimension-changing move is proposed 30% of times and we know the hyperparameters for the parameters of the priors.

Step 5. Proposal parameter specification

Set the parameter values for all the proposal distributions $q(\cdot|\cdot)$ to:

- an update in the cut point parameters $\{\mu_k\}$: $\sigma^2_{\mu_p} = 0.3$,
- an update in the row cluster parameters $\{\alpha_r\}$: $\sigma^2_{\alpha_p} = 0.3$,
- an update in the column parameters $\{\beta_j\}$: $\sigma^2_{\beta_p} = 0.3$,
- an update in the row group membership prob. parameters $\{\pi_r\}$: $\sigma^2_{\pi_p} = 0.3$.

Step 6. Generate replicate datasets

For each replicate $h \in \{1, \ldots, H\}$ and each chain $s \in \{1, \ldots, S\}$:

- For each row $i = 1, \ldots, n$, generate row membership as an indicator vector

$$\mathbf{z}_i^{hs} = \left(Z_{i1}^{hs}, ..., Z_{iR}^{hs}\right) \sim \text{Multinomial}\left(1; \{\pi_r\}\right).$$

- For each column $j = 1, \ldots, m$ within each row $i = 1, \ldots, n$, generate the response ordinal variable

$$y_{ij}^{hs} | \mathbf{z}_i^{hs} = \boldsymbol{\delta}_r \sim \text{Stereotype}\left(\{\eta_{krj}\}_{k=1}^{q}\right).$$

  Here $\boldsymbol{\delta}_r$ is an indicator vector of length R, with 1 at location r and zero elsewhere. This implies that

$$\log\left(\frac{\mathrm{P}\left[y_{ij}^{hs} = k \mid \mathbf{z}_i^{hs} = \boldsymbol{\delta}_r\right]}{\mathrm{P}\left[y_{ij}^{hs} = 1 \mid \mathbf{z}_i^{hs} = \boldsymbol{\delta}_r\right]}\right) = \eta_{krj}.$$

There are $HS = 100 \times 10 = 1000$ possible combinations of replicate and chain: hs.

Step 7. Fit models. Run the RJMCMC sampler

- We run the RJMCMC sampler for the combination dataset hs and we obtain the chain RJ^{hs}.

  This means running the sampler $HST = 100 \times 10 \times 220000$ iterations in all.

  At iteration t ($t = 1, \ldots, \mathrm{nstore}$) we obtain the estimated parameter vector $\widehat{\Omega}^{hs}_{(t)}$ for the combination dataset hs consisting of free parameters:

  $$\{\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{R-1}\}, \{\widehat{\beta}_1, \ldots, \widehat{\beta}_{m-1}\}, \{\widehat{\pi}_1, \ldots, \widehat{\pi}_{R-1}\}, \{\widehat{\mu}_2, \ldots, \widehat{\mu}_q\}, \text{ and } \{\widehat{\phi}_2, \ldots, \widehat{\phi}_{q-1}\}.$$

  The dimension of $\widehat{\Omega}^{hs}_{(t)}$ is variable depending on which dimension (in terms of number of rows R^{RJ}) the sampler is exploring. So, the chain s in the h^{th} replicate RJ^{hs} can be divided into a set of $R^{\mathrm{RJ}}_{\max} - R^{\mathrm{RJ}}_{\min} + 1$ subchains of fixed dimension ($R^{\mathrm{RJ}}_{\min} = 1$, $R^{\mathrm{RJ}}_{\max} = 10$)

  $$\mathrm{RJ}^{hs} = \mathrm{RJ}^{hs}\left(R^{\mathrm{RJ}}_{\min}\right) \cup \mathrm{RJ}^{hs}\left(R^{\mathrm{RJ}}_{\min+1}\right) \cup \cdots \cup \mathrm{RJ}^{hs}\left(R^{\mathrm{RJ}}_{\max}\right),$$

  where RJ^{hs} are the chains with r clusters, $r = R_{\min}, \ldots, R_{\max}$.

- Return the values $\left\{\widehat{\Omega}^{hs}_{(1)}, \widehat{\Omega}^{hs}_{(2)}, \ldots, \widehat{\Omega}^{hs}_{(\mathrm{nstore})}\right\}$ for $h = 1, \ldots, H$ and $s = 1, \ldots, S$.

- Test whether the convergence has been achieved. If not, increase nstore and return to step 7.

- Test whether the label-switching problem is present in the posterior distributions of $\{\alpha_r\}$ and $\{\pi_r\}$. If so, perform the procedure described in Section 7.2.5.

Step 8. Obtain overall results

- Merge the HS combinations of replicate and chain into one chain:

  $$\mathrm{RJ}^{\mathrm{overall}} = \bigcup_{h=1}^{H} \bigcup_{s=1}^{S} \mathrm{RJ}^{hs}.$$

- Graph a bar diagram depicting the number of iterations at which the chain $\mathrm{RJ}^{\mathrm{overall}}$ visits the model with R^{RJ} row clusters ($R^{\mathrm{RJ}}_{\mathrm{min}} \leq R^{\mathrm{RJ}} \leq R^{\mathrm{RJ}}_{\mathrm{max}}$).

- For each chain RJ^{hs}, calculate the mode of the number of row clusters visited by the chain:

$$\widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}} = \underset{R^{\mathrm{RJ}}_{\mathrm{min}} \leq R^{\mathrm{RJ}} \leq R^{\mathrm{RJ}}_{\mathrm{max}}}{\mathrm{argmax}} \dim\left(\mathrm{RJ}^{hs}\left(R^{\mathrm{RJ}}\right)\right)$$

  and select the section of the chain where the number of row clusters is the same as the mode:

$$\mathrm{RJ}^{hs}_{\mathrm{mode}} = \mathrm{RJ}^{hs}\left(R^{\mathrm{RJ}} = \widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}\right).$$

  After this point, we will have HS modes.

- The proportion of times across the HS possible combinations of replicate and chain where the mode $\widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}$ agrees with the true value R is of primary interest:

$$P_R\left(R, \widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}\right) = \frac{1}{HS}\sum_{h=1}^{H}\sum_{s=1}^{S} I(R = \widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}),$$

- We also compute a measure of *spread* as the proportion of chains RJ^{hs} across the HS possible chains where the mode $\widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}$ falls into the range from $R-1$ to $R+1$ row clusters. Thus, we can show the spread from the previous expression $P_R(\cdot,\cdot)$ as the interval:

$$\left[P_R\left(\max\{\mathrm{R}^{\mathrm{RJ}}_{\mathrm{min}}, \mathrm{R}-1\}, \widehat{\mathrm{R}}^{\mathrm{RJ}}_{\mathrm{mode}}\right), P_R\left(\min\{\mathrm{R}+1, \mathrm{R}^{\mathrm{RJ}}_{\mathrm{max}}\}, \widehat{\mathrm{R}}^{\mathrm{RJ}}_{\mathrm{mode}}\right)\right].$$

- We are also interested in the proportion of times the 95% HPD region includes the true value of the parameter across the HS possible chains for the

following parameters (those not dimensional-dependent):

$$P_{\mu_k} = \frac{1}{HS}\sum_{h=1}^{H}\sum_{s=1}^{S} I(\mu_k^{\text{true}} \in \text{HPD}_{\mu_k}^{hs}) \qquad k = 2, \ldots, q,$$

$$P_{\phi_k} = \frac{1}{HS}\sum_{h=1}^{H}\sum_{s=1}^{S} I(\phi_k^{\text{true}} \in \text{HPD}_{\phi_k}^{hs}) \qquad k = 2, \ldots, q-1, \text{ and}$$

$$P_{\beta_j} = \frac{1}{HS}\sum_{h=1}^{H}\sum_{s=1}^{S} I(\beta_j^{\text{true}} \in \text{HPD}_{\beta_j}^{hs}) \qquad j = 1, \ldots, m-1.$$

and the same for parameters $\{\alpha_r\}$ and $\{\pi_r\}$ (those dimensional-dependent) for the set of chains where the number of row clusters is $\widehat{R}^{\text{RJ}}_{\text{mode}}$:

$$P_{\alpha_r} = \frac{1}{HS}\sum_{h=1}^{H}\sum_{s=1}^{S} I(\alpha_r^{\text{true}} \in \text{HPD}_{\alpha_r}^{hs}) \qquad r = 1, \ldots, R-1, \text{ and}$$

$$P_{\pi_r} = \frac{1}{HS}\sum_{h=1}^{H}\sum_{s=1}^{S} I(\pi_r^{\text{true}} \in \text{HPD}_{\pi_r}^{hs}) \qquad r = 1, \ldots, R-1.$$

Step 9. Summarising results for each chain

- For chain RJ^{hs},
  - We summarise the chain RJ^{hs} computing the mean, median, interquartile range, standard deviation, time series standard error, highest posterior density interval (HPD) and maximum a posteriori estimator (MAP) for each element of the parameter vectors $\{\widehat{\Omega}^{hs}_{(1)}, \ldots, \widehat{\Omega}^{hs}_{(\text{nstore})}\}$.
  - Calculate the model averaged estimates computing the marginal posterior distributions conditional to the data $\boldsymbol{Y}$ and any possible model with R^{RJ} row clusters ($R^{\text{RJ}}_{\text{min}} \leq R^{\text{RJ}} \leq R^{\text{RJ}}_{\text{max}}$) for the following free parameters:

    $p(\mu_2|\boldsymbol{Y}, R \in \{R^{\text{RJ}}_{\text{min}}, \ldots, R^{\text{RJ}}_{\text{max}}\}), \ldots, p(\mu_q|\boldsymbol{Y}, R \in \{R^{\text{RJ}}_{\text{min}}, \ldots, R^{\text{RJ}}_{\text{max}}\}),$
    $p(\phi_2|\boldsymbol{Y}, R \in \{R^{\text{RJ}}_{\text{min}}, \ldots, R^{\text{RJ}}_{\text{max}}\}), \ldots, p(\phi_{q-1}|\boldsymbol{Y}, R \in \{R^{\text{RJ}}_{\text{min}}, \ldots, R^{\text{RJ}}_{\text{max}}\}),$ and
    $p(\beta_1|\boldsymbol{Y}, R \in \{R^{\text{RJ}}_{\text{min}}, \ldots, R^{\text{RJ}}_{\text{max}}\}), \ldots, p(\beta_{m-1}|\boldsymbol{Y}, R \in \{R^{\text{RJ}}_{\text{min}}, \ldots, R^{\text{RJ}}_{\text{max}}\}).$

    In addition, we also calculate model specific estimates computing the marginal posterior distributions conditional on the data $\boldsymbol{Y}$ and the

model with the mode of number of row clusters visited by the chain for the following free parameters:

$$p(\alpha_1|\boldsymbol{Y}, \widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}), \ldots, p(\alpha_{R^{\mathrm{RJ}}-1}|\boldsymbol{Y}, \widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}), \text{ and}$$
$$p(\pi_1|\boldsymbol{Y}, \widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}), \ldots, p(\pi_{R^{\mathrm{RJ}}-1}|\boldsymbol{Y}, \widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}).$$

This simulation study is for the one-dimensional clustering case and is illustrated with the row clustering version. The column clustering version is essentially the same, but replacing parameters related to rows with the equivalent column parameters. The biclustering version may be a future direction to explore.

8.6 Results

In this section, the use of the RJMCMC sampler to estimate the stereotype model parameters for our one-dimensional likelihood-based clustering approach is illustrated. We show the results of the simulation study described above (Section 8.5). Additionally, the results of this estimation approach for two real-life dataset examples are demonstrated.

Simulation Study Results

We summarise the simulation study results by computing the mean, median, interquartile range (IQR), standard deviation (SD), time-series standard error, highest posterior density interval (HPD) and maximum a posteriori estimator (MAP) for each parameter vector Ω^{hs} ($h = 1, \ldots, 100$ and $s = 1, \ldots, 10$).

Tables 8.3 and 8.4 show the results for the row clustering model for different number of row clusters R (from $R = 2$ to $R = 6$). In each case the tables show the mean of those statistical measures over the $HS = 100 \times 10 = 1000$ possible combinations of replicate and chain. The MAP estimators of all the parameters are close to their true values and as expected the 95% HPD credible interval includes the true parameter values in all the cases. Additionally, Figure 8.2 shows the $HS = 100 \times 10 = 1000$ separate MAP estimators of all the parameters taken in pairs and plotted against each other for the row clustering model with $R = 2$ row clusters. The red diamond point represents the true value of the parameter. The MAP estimators are around the true value of the parameter in all the scat-

ter plots. Note that all the MAP estimators on the top left plot which relates to the comparison between $\widehat{\phi}_2$ vs. $\widehat{\phi}_3$ show the ordering constraint $\widehat{\phi}_2 < \widehat{\phi}_3$. Also note that top right plot comparing the ($\widehat{\mu}_2$, $\widehat{\mu}_3$) pair shows a positive relationship. This is because the parameters $\{\mu_k\}$ determine the proportion of subjects per response category and therefore when the relative size of one high ordinal category increases then the adjacent categories also tend to increase. Figures J.1-J.4 in Appendix J show similar results for the row clustering model with $R = 3, \ldots, 6$ row clusters respectively.

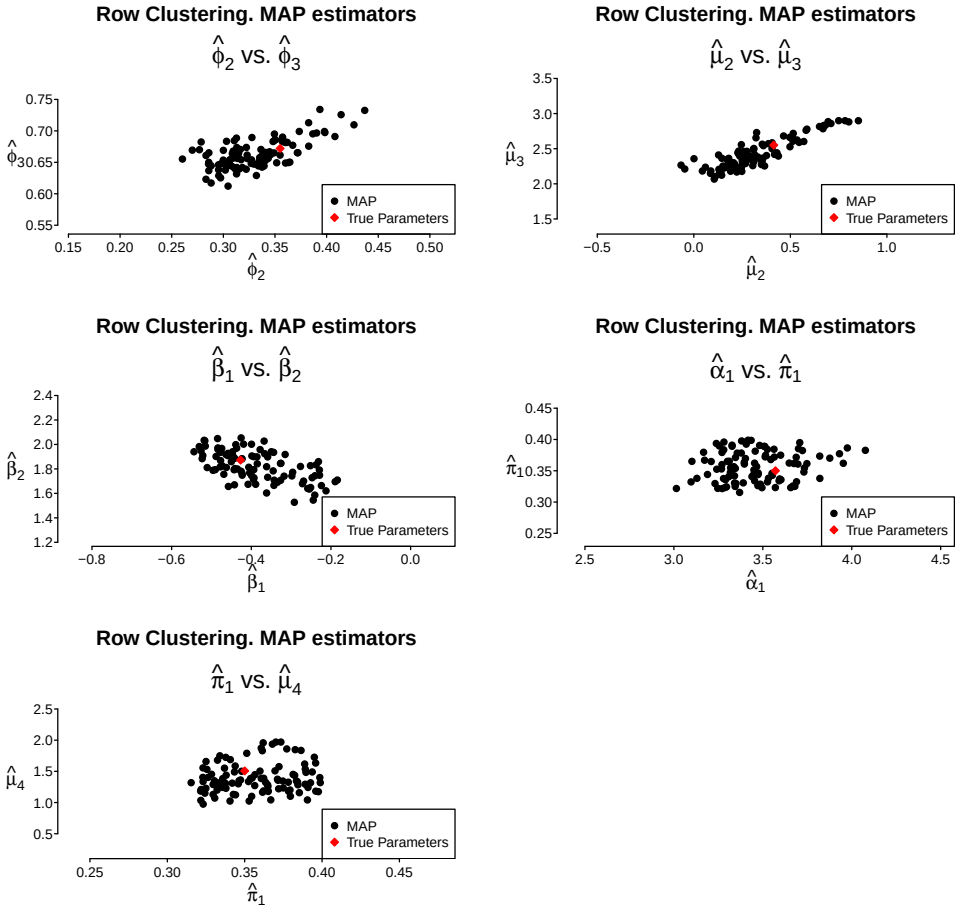

Figure 8.2: ***RJMCMC simulation study R=2****: Scatter plots depicting the maximum a posteriori estimator (MAP) across all the replicates ($H = 100$) and chains ($S = 10$) for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ with $R = 2$ row clusters. The sample size for each chain and replica is $n = 1000$, the number of categories is $q = 4$, and the number of columns is $m = 3$. The black points are the MAP estimators and the red diamond point represents the true value of the parameter.*

For each number of row clusters tested, we have merged the HS possible com-

Table 8.3: **RJMCMC simulation study**: Summary statistics for estimated parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. Mean, median, interquartile range (IQR), standard deviation (SD), time-series standard error, 95% highest posterior density interval (HPD), and maximum a posteriori estimator (MAP) for different number of row clusters R ($R = 2$, $R = 3$ and $R = 4$) are shown. The sample size is $n = 1000$, the number of categories is $q = 4$ and the number of columns is $m = 3$.

R	True parameters	RJMCMC							
		Mean	Median	IQR	S.D.	Time-series S.E.	HPD 95% lower	HPD 95% upper	MAP
2	$\mu_2 = 0.414$	0.300	0.055	0.579	0.240	0.014	-0.153	0.776	0.321
	$\mu_3 = 2.551$	2.421	1.976	1.118	0.245	0.015	1.968	2.907	2.428
	$\mu_4 = 1.507$	1.380	0.725	1.668	0.275	0.017	0.868	1.928	1.389
	$\phi_2 = 0.355$	0.329	0.328	0.066	0.049	0.002	0.236	0.424	0.332
	$\phi_3 = 0.672$	0.664	0.665	0.038	0.028	0.001	0.611	0.718	0.664
	$\beta_1 = -0.427$	-0.387	-0.392	0.137	0.100	0.003	-0.549	-0.163	-0.387
	$\beta_2 = 1.872$	1.818	1.837	0.160	0.115	0.004	1.595	2.041	1.824
	$\alpha_1 = 3.571$	3.458	3.445	0.303	0.235	0.014	3.027	3.915	3.460
	$\pi_1 = 0.350$	0.319	0.358	0.029	0.025	0.001	0.288	0.368	0.338
3	$\mu_2 = 0.414$	0.466	0.394	0.375	0.278	0.016	-0.039	1.022	0.460
	$\mu_3 = 2.551$	2.649	2.528	0.570	0.407	0.026	1.925	3.471	2.605
	$\mu_4 = 1.507$	1.671	1.508	0.812	0.572	0.037	0.656	2.823	1.602
	$\phi_2 = 0.355$	0.344	0.347	0.086	0.064	0.002	0.218	0.467	0.350
	$\phi_3 = 0.672$	0.674	0.675	0.047	0.035	0.001	0.607	0.741	0.682
	$\beta_1 = -0.427$	-0.449	-0.450	0.148	0.109	0.003	-0.661	-0.238	-0.459
	$\beta_2 = 1.872$	1.860	1.863	0.194	0.144	0.004	1.585	2.141	1.870
	$\alpha_1 = 3.571$	3.254	3.253	0.853	0.680	0.024	1.892	4.566	3.624
	$\alpha_2 = -0.919$	-0.710	-0.798	0.724	0.683	0.029	-1.699	0.737	-0.757
	$\pi_1 = 0.200$	0.169	0.168	0.091	0.067	0.002	0.045	0.288	0.135
	$\pi_2 = 0.500$	0.529	0.531	0.172	0.117	0.005	0.319	0.743	0.581
4	$\mu_2 = 0.414$	0.253	0.234	0.407	0.311	0.019	-0.335	0.855	0.346
	$\mu_3 = 2.551$	2.306	2.292	0.637	0.490	0.034	1.350	3.216	2.474
	$\mu_4 = 1.507$	1.194	1.182	0.937	0.725	0.050	-0.217	2.523	1.434
	$\phi_2 = 0.355$	0.358	0.358	0.071	0.053	0.001	0.254	0.461	0.362
	$\phi_3 = 0.672$	0.661	0.661	0.042	0.032	0.001	0.600	0.722	0.663
	$\beta_1 = -0.427$	-0.442	-0.442	0.147	0.109	0.003	-0.654	-0.232	-0.444
	$\beta_2 = 1.872$	1.783	1.782	0.176	0.130	0.003	1.531	2.036	1.804
	$\alpha_1 = 3.571$	3.726	3.726	0.969	0.750	0.023	2.257	5.169	4.024
	$\alpha_2 = -0.919$	-0.944	-0.871	1.170	0.829	0.030	-2.611	0.386	-0.788
	$\alpha_3 = 1.228$	1.191	1.119	1.052	0.790	0.027	-0.069	2.811	0.933
	$\pi_1 = 0.250$	0.251	0.251	0.054	0.049	0.001	0.150	0.355	0.253
	$\pi_2 = 0.320$	0.341	0.339	0.082	0.074	0.002	0.170	0.481	0.343
	$\pi_3 = 0.150$	0.123	0.128	0.073	0.049	0.001	0.025	0.209	0.129

Table 8.4: **RJMCMC simulation study**: Summary statistics for estimated parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. Mean, median, interquartile range (IQR), standard deviation (SD), time-series standard error, 95% highest posterior density interval (HPD), and maximum a posteriori estimator (MAP) for different number of row clusters R ($R = 5$ and $R = 6$) are shown. The sample size is $n = 1000$, the number of categories is $q = 4$ and the number of columns is $m = 3$.

R	True parameters	RJMCMC							
		Mean	Median	IQR	S.D.	Time-series S.E.	HPD 95% lower	HPD 95% upper	MAP
5	$\mu_2 = 0.414$	0.183	0.226	0.540	0.361	0.030	-0.510	0.880	0.174
	$\mu_3 = 2.551$	2.091	2.217	0.934	0.604	0.055	0.910	3.218	2.057
	$\mu_4 = 1.507$	0.792	1.004	1.383	0.886	0.081	-0.936	2.437	0.709
	$\phi_2 = 0.355$	0.358	0.355	0.053	0.039	0.002	0.282	0.434	0.358
	$\phi_3 = 0.672$	0.674	0.674	0.039	0.029	0.001	0.618	0.729	0.673
	$\beta_1 = -0.427$	-0.392	-0.391	0.170	0.126	0.005	-0.634	-0.147	-0.409
	$\beta_2 = 1.872$	1.859	1.854	0.196	0.145	0.006	1.579	2.138	1.901
	$\alpha_1 = 2.571$	2.779	2.817	1.595	1.116	0.050	0.604	4.800	2.853
	$\alpha_2 = -2.919$	-2.726	-2.676	1.202	0.924	0.045	-4.618	-0.992	-2.722
	$\alpha_3 = 1.528$	0.665	0.320	1.517	1.613	0.078	-1.671	4.550	1.425
	$\alpha_4 = 6.012$	6.091	6.094	1.251	1.187	0.055	4.331	7.978	6.455
	$\pi_1 = 0.200$	0.200	0.200	0.028	0.024	0.001	0.152	0.247	0.202
	$\pi_2 = 0.200$	0.197	0.195	0.070	0.044	0.002	0.121	0.276	0.195
	$\pi_3 = 0.200$	0.174	0.158	0.100	0.063	0.003	0.077	0.290	0.167
	$\pi_4 = 0.200$	0.161	0.126	0.153	0.096	0.004	0.026	0.329	0.140
6	$\mu_2 = 0.414$	0.344	0.314	0.449	0.337	0.022	-0.301	1.003	0.341
	$\mu_3 = 2.551$	2.378	2.345	0.755	0.559	0.040	1.297	3.457	2.333
	$\mu_4 = 1.507$	1.223	1.187	1.117	0.819	0.058	-0.378	2.773	1.134
	$\phi_2 = 0.355$	0.353	0.353	0.057	0.042	0.001	0.272	0.436	0.358
	$\phi_3 = 0.672$	0.673	0.673	0.041	0.031	0.001	0.614	0.733	0.679
	$\beta_1 = -0.427$	-0.430	-0.429	0.164	0.122	0.003	-0.666	-0.194	-0.438
	$\beta_2 = 1.872$	1.841	1.838	0.194	0.144	0.004	1.563	2.122	1.864
	$\alpha_1 = 2.571$	2.564	2.545	1.322	0.953	0.027	0.743	4.396	2.369
	$\alpha_2 = -2.919$	-2.719	-2.670	1.309	0.945	0.032	-4.587	-0.989	-2.843
	$\alpha_3 = 1.528$	0.932	0.769	1.237	1.172	0.034	-1.040	3.686	1.514
	$\alpha_4 = 6.012$	5.819	5.773	1.249	0.968	0.036	3.992	7.720	6.064
	$\alpha_5 = -0.512$	-0.227	-0.421	1.248	1.872	0.058	-3.049	5.055	-0.264
	$\pi_1 = 0.170$	0.169	0.169	0.024	0.021	0.001	0.125	0.212	0.174
	$\pi_2 = 0.170$	0.162	0.158	0.058	0.038	0.001	0.094	0.232	0.152
	$\pi_3 = 0.170$	0.163	0.152	0.097	0.057	0.002	0.069	0.262	0.169
	$\pi_4 = 0.170$	0.140	0.114	0.123	0.072	0.002	0.041	0.278	0.131
	$\pi_5 = 0.170$	0.132	0.088	0.181	0.097	0.003	0.012	0.299	0.124

binations of replicate and chain into one single chain $\mathrm{RJ}^{\mathrm{overall}}$ in order to show the results summarised globally. In each case, the resultant merged chain $\mathrm{RJ}^{\mathrm{overall}}$ is thinned to 6000 iterations to ease its depiction. Figures 8.3 and 8.4 show the marginal posterior distribution and trace plot for all the parameters for $R = 2$ number of row clusters respectively. The expected values of the posterior distribution are very close to the true values (green vertical lines) and the 95% HPD credible interval includes the true parameter values in all the cases. The trace plots on Figure 8.4 show a good mixing on all the parameters. The related graphics for $R = 3, \ldots, 6$ are shown in Figures J.5-J.12 in Appendix J showing similar satisfactory results.

Figure 8.5 shows a bar diagram depicting the number of iterations at which the merged chain $\mathrm{RJ}^{\mathrm{overall}}$ visits the model with $\widehat{R}^{\mathrm{RJ}}$ row clusters ($1 \leq \widehat{R}^{\mathrm{RJ}} \leq 10$) for each number of true row clusters R tested (from $R_{\min} = 2$ to $R_{\max} = 6$). The mode $\widehat{R}^{\mathrm{RJ}}_{\mathrm{mode}}$ for each plot always corresponds to the related true value R showing that the model related to this cluster dimension was the most visited by the RJMCMC sampler. Additionally, Table 8.5 shows the posterior probabilities by all the sub-models. In all the cases, the RJMCMC sampler does spend more time in sub-models in the vicinity of the true model with R clusters. The posterior

Table 8.5: **RJMCMC simulation study**: Posterior probabilities by sub-model for row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ model with $R = 2, \ldots, 6$. The number of clusters visited by the RJMCMC sampler are $R^{\mathrm{RJ}} = 1 \ldots 10$. The highest probability for each true number of cluster R (depicted in the columns) is shown in boldface.

$\boldsymbol{R^{\mathrm{RJ}}}$	**Posterior Probabilities (%)**				
	$\boldsymbol{R=2}$	$\boldsymbol{R=3}$	$\boldsymbol{R=4}$	$\boldsymbol{R=5}$	$\boldsymbol{R=6}$
1	0.11	0.17	0.05	0.05	0.01
2	**58.42**	31.48	7.17	1.45	0.08
3	25.70	**51.40**	26.52	3.77	3.51
4	12.31	12.67	**41.92**	20.81	8.40
5	3.02	3.18	17.75	**41.26**	18.83
6	0.41	0.82	5.20	22.77	**31.27**
7	0.01	0.22	1.17	7.69	27.55
8	0.01	0.04	0.19	1.81	9.13
9	0.00	0.01	0.02	0.35	1.16
10	0.00	0.01	0.01	0.04	0.06

probabilities of these models are typically between 30% and 60% showing that

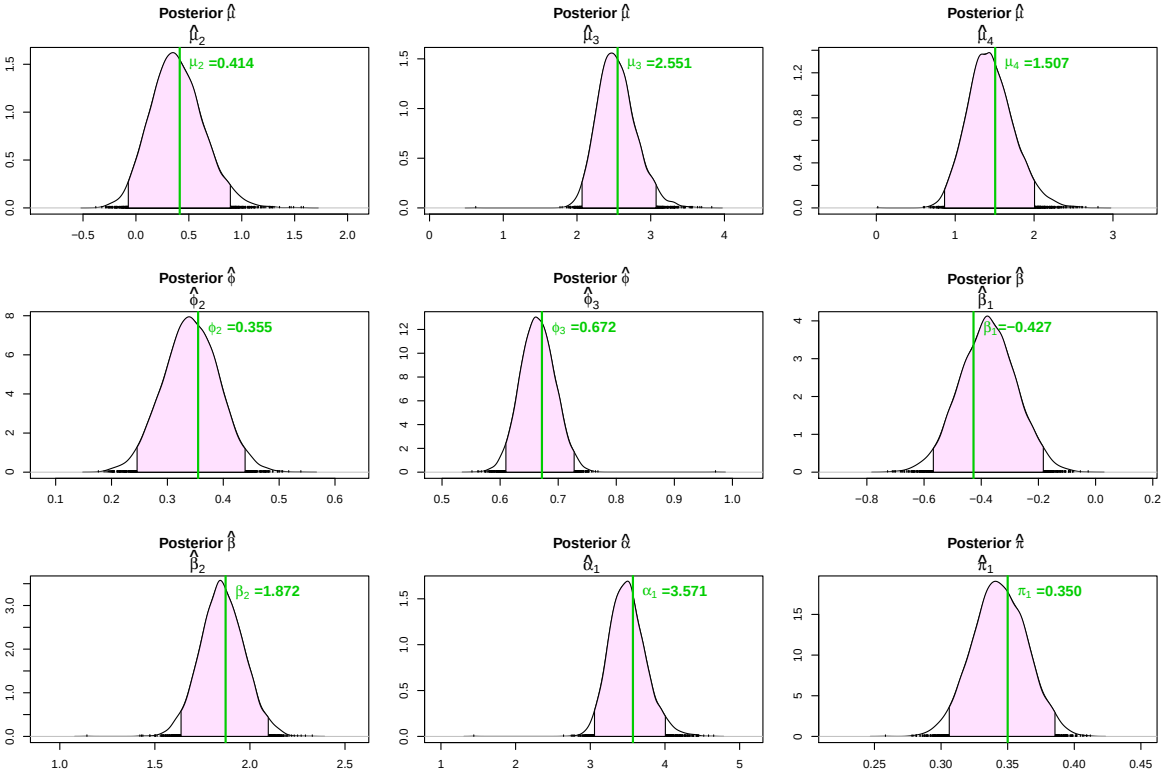

Figure 8.3: ***RJMCMC simulation study R=2****: Density plot of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$*. The sample size is* $n = 1000$*, the number of categories is* $q = 4$*, the number of columns is* $m = 3$ *and the number of row clusters are* $R = 2$*. The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 10$*). The green vertical lines are the true parameter value and 95% HPD credible intervals are shown with shading area.*

the number of clusters is not easy to identify by the sampler particularly when the true number of clusters is $R \geq 4$. However, most of the posterior mass is on models with R in the neighborhood ($\widehat{R}^{\text{RJ}}_{\text{mode}} \pm 1$) of the true number of clusters (typically more than 75%). Table 8.6 summarise this results showing the proportion of times across the HS possible combinations where the mode $\widehat{R}^{\text{RJ}}_{\text{mode}}$ agrees with the true value R. We also show on this table a spread measure computed as the proportion of chains where the mode $\widehat{R}^{\text{RJ}}_{\text{mode}}$ falls into the range from $\widehat{R}^{\text{RJ}}_{\text{mode}} - 1$ to $\widehat{R}^{\text{RJ}}_{\text{mode}} + 1$. All the proportions are greater than 77% and best results are for the scenario when $R = 3$ where 95.6% of times the spread is covering the true

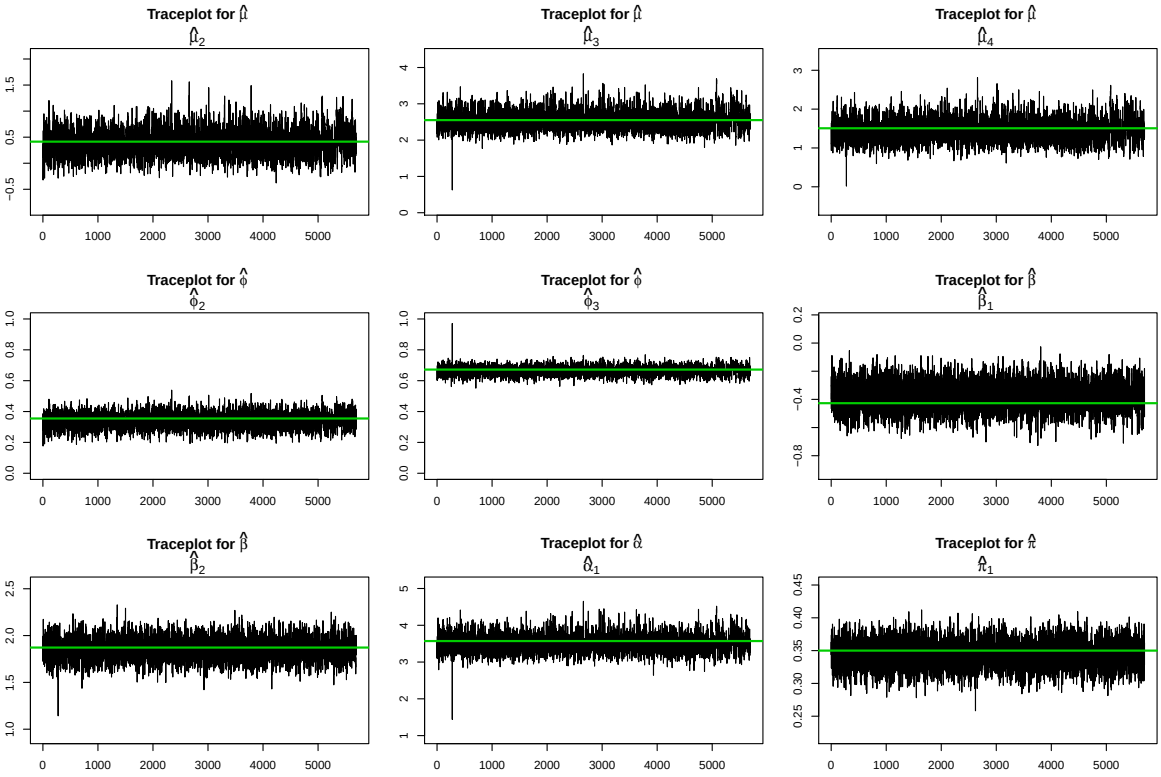

Figure 8.4: ***RJMCMC simulation study R=2****: Trace plot of the parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The sample size is $n = 1000$, the number of categories is $q = 4$, the number of columns is $m = 3$ and the number of row clusters are $R = 2$. The plots depict the results of the RJMCMC sampler for a sample of 6000 iterations over all the replicates ($H = 100$) and chains ($S = 10$). The green horizontal lines are the true parameter value.*

Table 8.6: **RJMCMC simulation study**: Proportion of times across the $HS = 100 \times 10 = 1000$ possible combinations of replicate and chain where the mode $\widehat{R}^{\text{RJ}}_{\text{mode}}$ agrees with the true value R (from $R = 2$ to $R = 6$) for the stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The posterior probabilities of the spread interval ($R_{\text{mode}} \pm 1$) are also shown.

True $\boldsymbol{R}$	$\boldsymbol{R = 2}$	$\boldsymbol{R = 3}$	$\boldsymbol{R = 4}$	$\boldsymbol{R = 5}$	$\boldsymbol{R = 6}$
Mode ($\widehat{R}^{\text{RJ}}_{\text{mode}}$)	58.4%	51.4%	41.9%	41.3%	31.3%
Spread ($\widehat{R}^{\text{RJ}}_{\text{mode}} \pm 1$)	84.2%	95.6%	86.2%	84.8%	77.7%

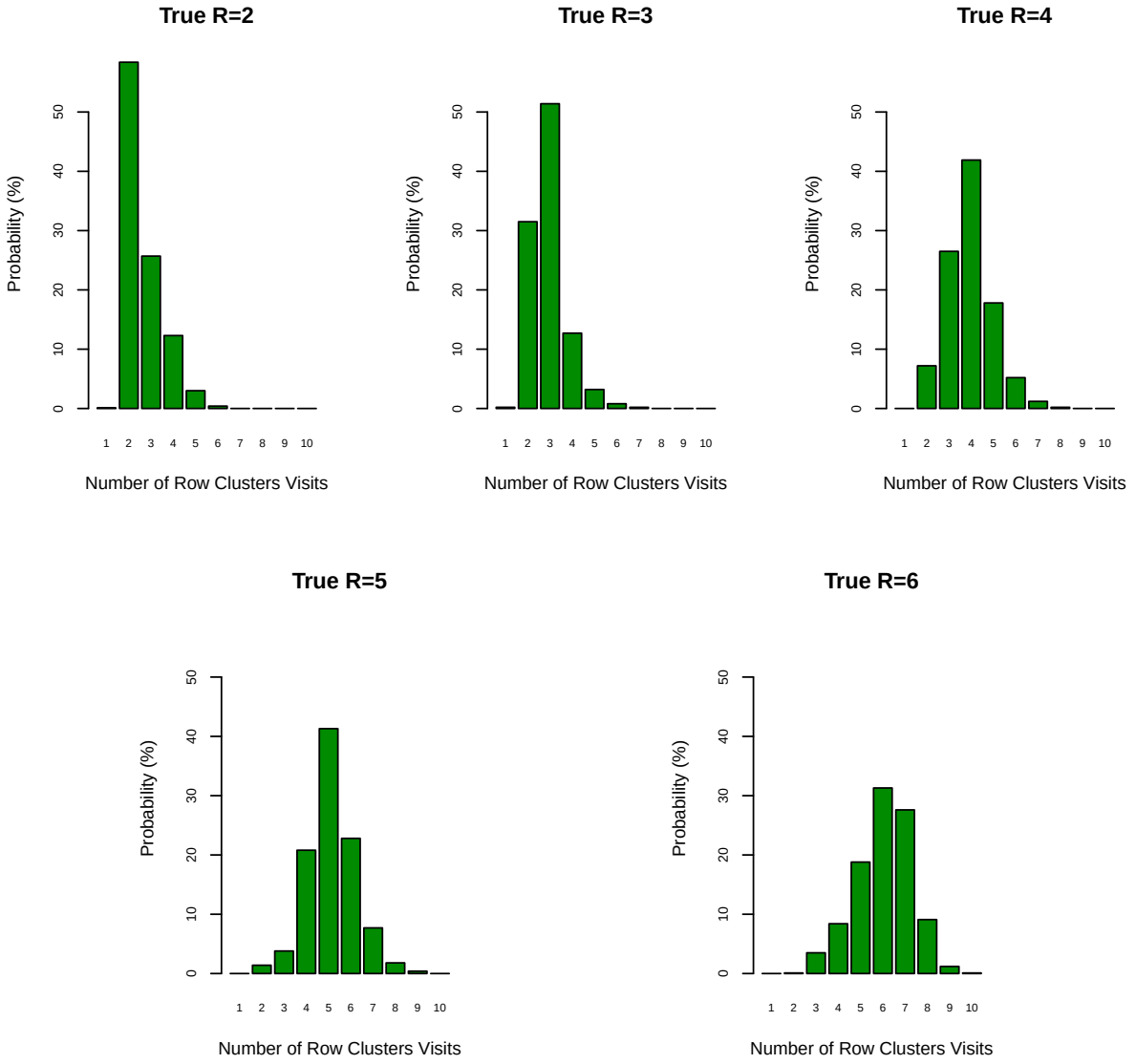

Figure 8.5: ***RJMCMC simulation study***: *Bar plots depicting the number of rows R visited by the RJMCMC sampler in the estimation procedure for row clustering model $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$. Each plot represents the scenario with a true number of row clusters R ($R = 2$ to $R = 6$). The y-axis limits (posterior probabilities) are the same to make plots comparable. The sample size is $n = 1000$, the number of categories is $q = 4$ and the number of columns is $m = 3$.*

number of clusters R.

The proportion of times across the HS possible chains where the 95% HPD region includes the true value of the parameters with fixed dimension ($\{\mu_k\}$,$\{\phi_k\}$,

and $\{\beta_k\}$) is shown in Table 8.7. Table J.1 in Appendix J shows the equivalent results for those parameters with variable dimension ($\{\alpha_r\}$ and $\{\pi_r\}$). In both cases, the proportion of times that true parameters are covered by the 95% HPD is 90% at least which is very satisfactory.

Table 8.7: **RJMCMC simulation study**: Proportion of times the 95% HPD region includes the true value of the fixed-dimensional parameters across the HS possible chains for the stereotype model including row clustering without interaction factors $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$.

True parameters	**Proportion of times within 95% HPD region**				
	$\boldsymbol{R=2}$	$\boldsymbol{R=3}$	$\boldsymbol{R=4}$	$\boldsymbol{R=5}$	$\boldsymbol{R=6}$
$\mu_2 = 0.414$	93%	90%	94%	94%	93%
$\mu_3 = 2.551$	92%	95%	93%	94%	91%
$\mu_4 = 1.507$	98%	96%	93%	94%	92%
$\phi_2 = 0.355$	97%	99%	90%	90%	97%
$\phi_3 = 0.672$	97%	100%	91%	99%	92%
$\beta_1 = -0.427$	96%	98%	97%	97%	93%
$\beta_2 = 1.872$	94%	90%	91%	90%	94%

In conclusion, the initial results described above for our RJMCMC sampler are encouraging in their ability to estimate parameters correctly. Nevertheless, a comprehensive test of the success of the estimation in challenging situations where it might be expected that estimation might be difficult would be required and is left as a possible future development.

Real-Life Data Results

In this section we use our RJMCMC sampler to estimate the parameters for two real-life data examples, which were introduced in Sections 4.2.1 and 4.2.3 respectively. The RJMCMC results are compared with those obtained from fitting our suite of models by running the EM algorithm and using AIC to do model comparison. Using the RJMCMC sampler as described in this thesis allows us not to have to make model choices among the possible models accessible to the sampler.

In both examples, we have applied the RJMCMC sampler to one-dimensional clustering described in Section 8.3. In the Example 1, we ran the sampler for the row clustering case, and we did the same for the column clustering version in the

Example 2. We did this in order to compare with the maximum likelihood estimator values (MLE) from the EM algorithm. Additionally, we assessed the convergence of the RJMCMC sampler by running five chains in parallel from widely dispersed starting points. As in the simulation study (Section 8.5), each chain was run for an initial burn-in period of 20000 iterations which are discarded. We then ran each chain for a further 200000 updates, storing only every 10th state (thinning). We used the methods of Castelloe and Zimmerman (2002) to assess the convergence of these chains and the results are given in Appendices I.2 and I.3. Finally, we merged the chains and the resulting chain was thinned which is summarised computing the mean, standard deviation, median, interquartile range, standard error, highest posterior density interval and maximum a posteriori estimator for the free parameter vector.

Example 1: Applied Statistics Course Feedback

The dimensions of the ordinal data matrix for the Applied Statistics course feedback data set are $n = 70$ rows (students) and $m = 10$ columns (questions) where each observation can take one of the three possible categories ($q = 3$). Figure 8.6 shows a bar diagram depicting the number of iterations at which the chain visits the model with R row clusters within a range from $R = 1$ to $R = 10$. The mode was $\widehat{R} = 3$ which coincides with the number of clusters in the row clustering model selected by AIC using the EM algorithm estimation approach (see Table 4.4 in Section 4.2.1). Table 8.8 shows the summary of the RJMCMC sampler results and they are compared with those MLE values from the EM algorithm. These results include estimates related to dimensional-independent parameters ($\{\mu_k\}$, $\{\phi_k\}$, and $\{\beta_j\}$) over all possible models visited by the sampler, and also includes estimates for parameters $\{\alpha_r\}$ and $\{\pi_r\}$ (those dimensional-dependent) from the set of submodels with fixed dimension $\widehat{R} = 3$. Moreover, Figure 8.7 depicts the marginal posterior distributions for all the parameters. The expected values of the posterior distribution are very close to the MLE values (blue vertical lines). The trace plots on Figure 8.8 show a good mixing in the sampling of all the parameters.

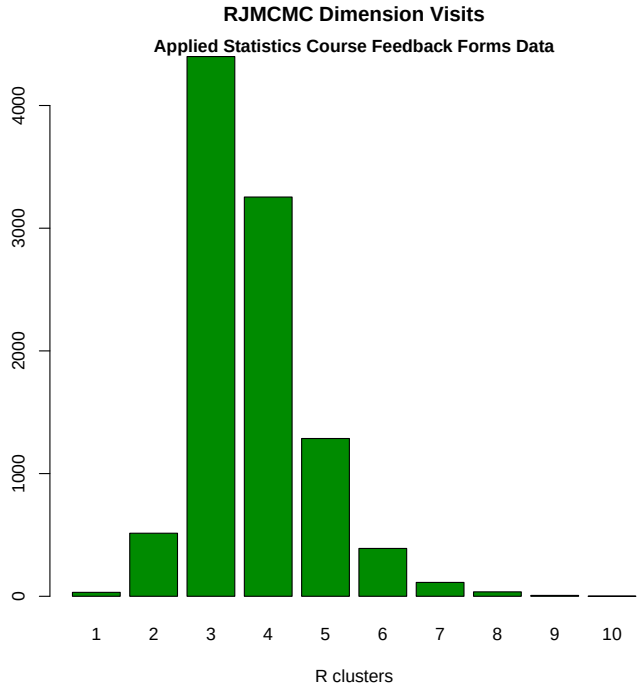

Figure 8.6: ***Applied Statistics course feedback forms data set:*** *Bar plot depicting the number of row clusters R visited by the RJMCMC sampler within the range $R = 1, \ldots, 10$ student clusters in the estimation procedure for row clustering model $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$.*

Example 2: Spider Data

The dimensions of the original count data matrix for the Spider data set are $n = 12$ spider species over $m = 28$ sites. Each observation was categorised in $q = 4$ ordinal responses as described in Section 4.2.3. Figure 8.9 shows a bar diagram depicting the number of iterations at which the chain visits the model with C site (column) clusters within a range from $C = 1$ to $C = 10$. The mode was $\widehat{C} = 3$ which agrees with the number of column clusters related to the best model using the EM algorithm approach and according to AIC (see Table C.12 in Appendix C.4). The comparison of the results for the two estimation approaches are described in Table 8.9 and Figure 8.10. The expected values on the marginal posterior distributions coincide with the MLE values (blue vertical lines). The trace plots on Figure 8.11 depict an acceptably good mixing in the sampling of all the parameters.

Table 8.8: **Applied Statistics course feedback forms data set:** Summary statistics for estimated parameters for stereotype model including row clustering model $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ with $R = 3$ row (student) clusters. The RJMCMC sampler results are compared with the MLE values from the EM algorithm.

Par.	EM algorithm		RJMCMC							
	Estim.	S.E.	Mean	Median	IQR	S.D.	Time-series S.E.	HPD 95% lower	HPD 95% upper	MAP
$\widehat{\mu}_2$	-0.306	0.140	-0.385	-0.339	0.361	0.339	0.010	-1.113	0.224	-0.122
$\widehat{\mu}_3$	-2.291	0.307	-2.271	-2.237	0.708	0.568	0.015	-3.516	-1.277	-1.905
$\widehat{\phi}_2$	0.541	0.195	0.553	0.549	0.093	0.070	0.002	0.422	0.693	0.515
$\widehat{\alpha}_1$	3.496	0.346	3.249	3.439	0.999	0.903	0.034	1.110	4.621	3.695
$\widehat{\alpha}_2$	-3.571	0.222	-3.869	-3.750	1.309	0.874	0.042	-5.590	-2.340	-3.530
$\widehat{\beta}_1$	-1.390	0.312	-1.365	-1.355	0.695	0.530	0.010	-2.482	-0.382	-1.419
$\widehat{\beta}_2$	-2.998	0.351	-3.063	-3.023	0.932	0.695	0.015	-4.457	-1.765	-3.440
$\widehat{\beta}_3$	-6.272	0.318	-6.318	-6.222	1.612	1.231	0.037	-8.782	-4.040	-6.381
$\widehat{\beta}_4$	0.300	0.437	0.246	0.251	0.609	0.457	0.009	-0.708	1.102	0.528
$\widehat{\beta}_5$	1.015	0.432	1.027	1.015	0.610	0.451	0.009	0.135	1.874	1.425
$\widehat{\beta}_6$	3.391	0.451	3.433	3.429	0.632	0.477	0.012	2.496	4.353	3.365
$\widehat{\beta}_7$	3.561	0.452	3.538	3.529	0.669	0.487	0.012	2.638	4.518	3.770
$\widehat{\beta}_8$	3.029	0.463	2.937	2.931	0.693	0.508	0.013	1.938	3.910	3.013
$\widehat{\beta}_9$	-1.601	0.332	-1.653	-1.612	0.736	0.565	0.011	-2.873	-0.644	-1.926
$\widehat{\pi}_1$	0.377	0.218	0.358	0.368	0.076	0.055	0.002	0.241	0.449	0.387
$\widehat{\pi}_2$	0.532	0.231	0.502	0.517	0.092	0.082	0.002	0.310	0.630	0.476

8.7 Discussion

The development of a RJMCMC sampler to apply to likelihood models based on the ordinal stereotype model and introducing fuzzy clustering via finite mixtures has been described in this chapter. The use of RJMCMC jointly with a label switching procedure allows for selection of the best model dimension while the sampler is estimating the model parameters. This is an advantage over a MCMC sampler such as the Metropolis-Hastings algorithm where the jump between models is not possible and therefore the different models have to be sampled independently. The reliability of using our one-dimensional models using RJMCMC has been tested with a simulation study and it was also compared with the MLE values for two real-data examples. Two of the drawbacks of this approach are that the sampler requires selection of suitable proposal distributions and the mixing might be slower than in fixed-dimensional MCMC samplers. Fu-

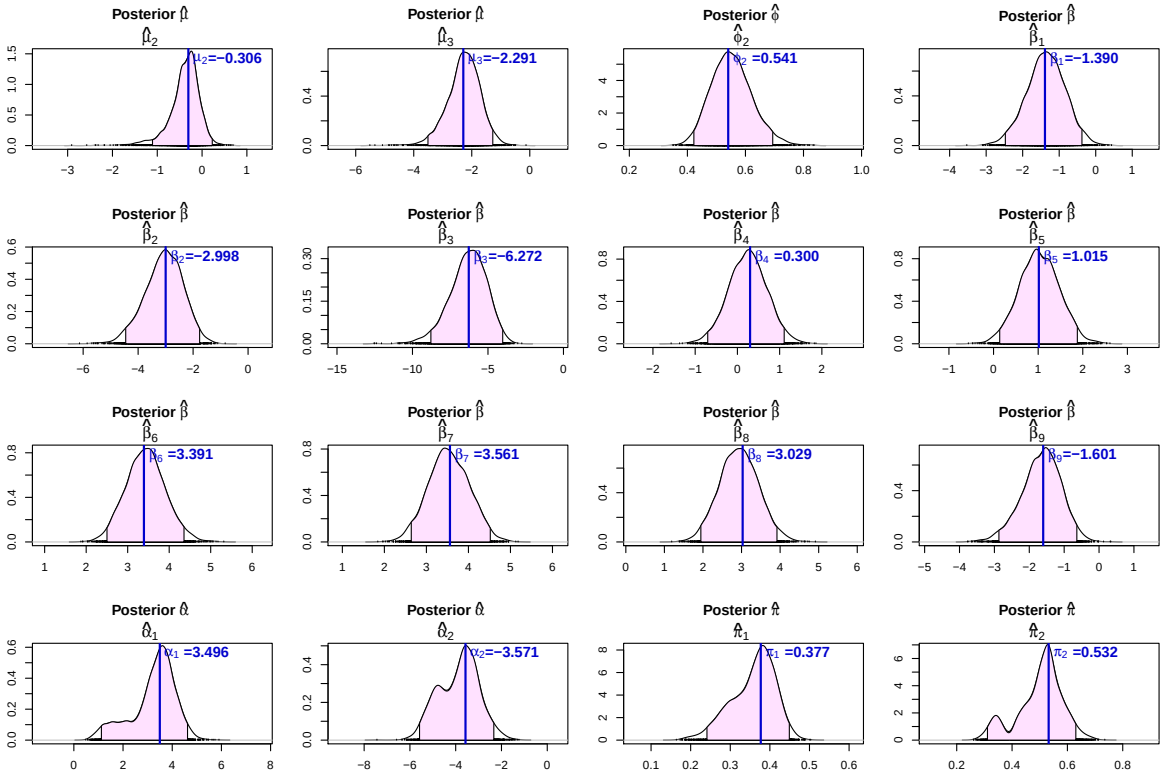

Figure 8.7: ***Applied Statistics course feedback forms data set:*** *Density plot depicting the marginal posterior distribution of the parameters for stereotype model including row clustering model* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ *with* $R = 3$ *row (student) clusters. The blue vertical lines are the MLE values. 95% HPD credible intervals are shown with shading area.*

ture developments in this area would include the development of an extra layer in our RJMCMC sampler allowing both jumps between different class families (i.e., between with and without interaction models from the same family) and jumps between one-dimensional and two-dimensional models (biclustering).

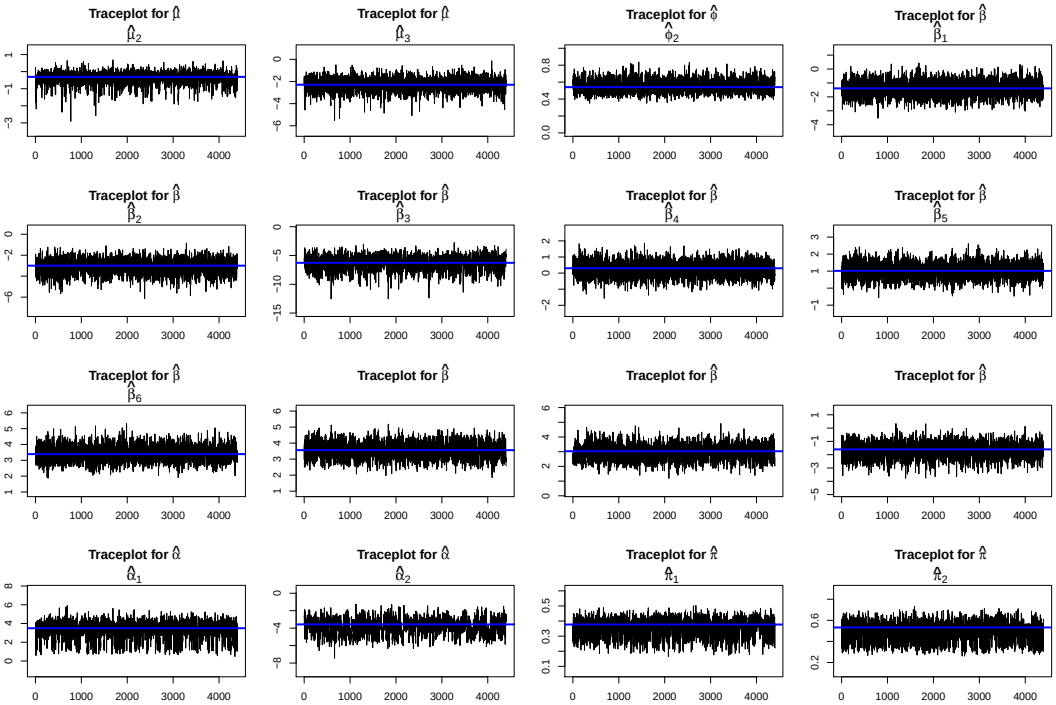

Figure 8.8: ***Applied Statistics course feedback forms data set:*** *Trace plot of the parameters for stereotype model including row clustering model* $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$ *with* $R = 3$ *row (student) clusters. The blue horizontal lines are the MLE values.*

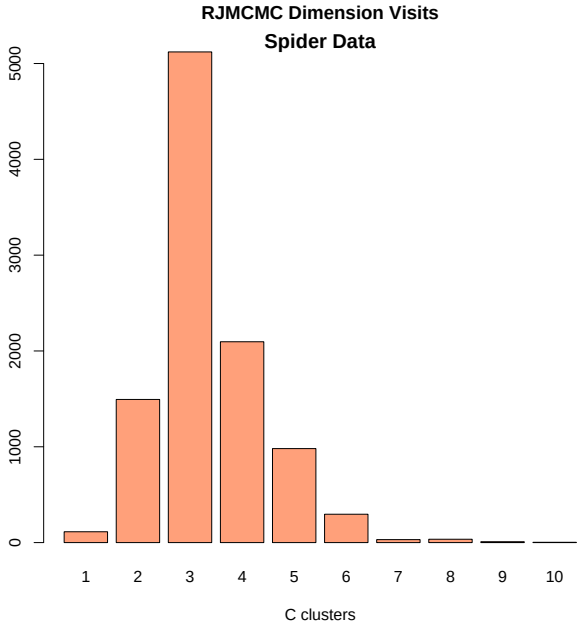

Figure 8.9: ***Spider data set:*** *Bar plot depicting the number of column clusters* C *visited by the RJMCMC sampler within the range* $C = 1, \ldots, 10$ *site clusters in the estimation procedure for column clustering model* $\boldsymbol{\mu_k + \phi_k(\alpha_i + \beta_c)}$.

Table 8.9: **Spider data set:** Summary statistics for estimated parameters for stereotype model including column clustering model $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_i} + \boldsymbol{\beta_c})$ with $C = 3$ column (site) clusters. The RJMCMC sampler results are compared with MLE values from the EM algorithm.

Par.	EM algorithm		RJMCMC							
	Estim.	S.E.	Mean	Median	IQR	S.D.	Time-series S.E.	HPD 95% lower	HPD 95% upper	MAP
$\widehat{\mu}_2$	0.131	0.195	0.173	0.179	0.312	0.239	0.006	-0.294	0.640	0.401
$\widehat{\mu}_3$	-0.812	0.162	-0.910	-0.885	0.680	0.558	0.019	-2.028	0.126	-1.371
$\widehat{\mu}_4$	-9.442	0.134	-9.974	-10.037	2.744	2.101	0.127	-14.609	-6.464	-12.743
$\widehat{\phi}_2$	0.397	0.139	0.398	0.399	0.122	0.090	0.003	0.227	0.578	0.364
$\widehat{\phi}_3$	0.889	0.119	0.889	0.890	0.061	0.045	0.001	0.811	0.979	0.915
$\widehat{\alpha}_1$	2.054	0.132	2.119	2.145	2.623	2.027	0.053	-1.769	6.167	2.016
$\widehat{\alpha}_2$	-1.470	0.098	-1.500	-1.435	2.671	2.079	0.054	-5.404	2.689	-1.142
$\widehat{\alpha}_3$	-0.005	0.097	-0.088	-0.093	2.940	2.203	0.054	-4.354	4.316	-2.074
$\widehat{\alpha}_4$	-3.816	0.125	-3.781	-3.660	3.801	2.899	0.076	-9.285	2.096	-6.147
$\widehat{\alpha}_5$	3.369	0.161	3.373	3.379	3.494	2.660	0.066	-1.813	8.585	4.688
$\widehat{\alpha}_6$	-2.245	0.082	-2.181	-2.053	3.215	2.466	0.072	-7.017	2.713	1.525
$\widehat{\alpha}_7$	-1.511	0.120	-1.434	-1.429	2.864	2.111	0.052	-5.816	2.479	-1.967
$\widehat{\alpha}_8$	0.524	0.105	0.549	0.571	2.527	1.901	0.058	-3.389	4.058	0.644
$\widehat{\alpha}_9$	-2.190	0.191	-2.265	-2.246	2.976	2.279	0.072	-6.678	2.309	-3.272
$\widehat{\alpha}_{10}$	-0.249	0.147	-0.315	-0.406	2.854	2.166	0.071	-4.373	4.128	-2.973
$\widehat{\alpha}_{11}$	5.744	0.105	5.791	5.765	2.841	2.123	0.058	1.893	10.176	8.407
$\widehat{\beta}_1$	-0.415	0.207	-0.346	-0.319	1.493	1.257	0.045	-2.853	1.979	0.687
$\widehat{\beta}_2$	-0.291	0.120	-0.212	-0.174	1.490	1.302	0.045	-2.790	2.117	1.587
$\widehat{\kappa}_1$	0.295	0.180	0.301	0.377	0.260	0.176	0.007	0.073	0.710	0.464
$\widehat{\kappa}_2$	0.348	0.194	0.381	0.402	0.256	0.180	0.009	0.090	0.747	0.470

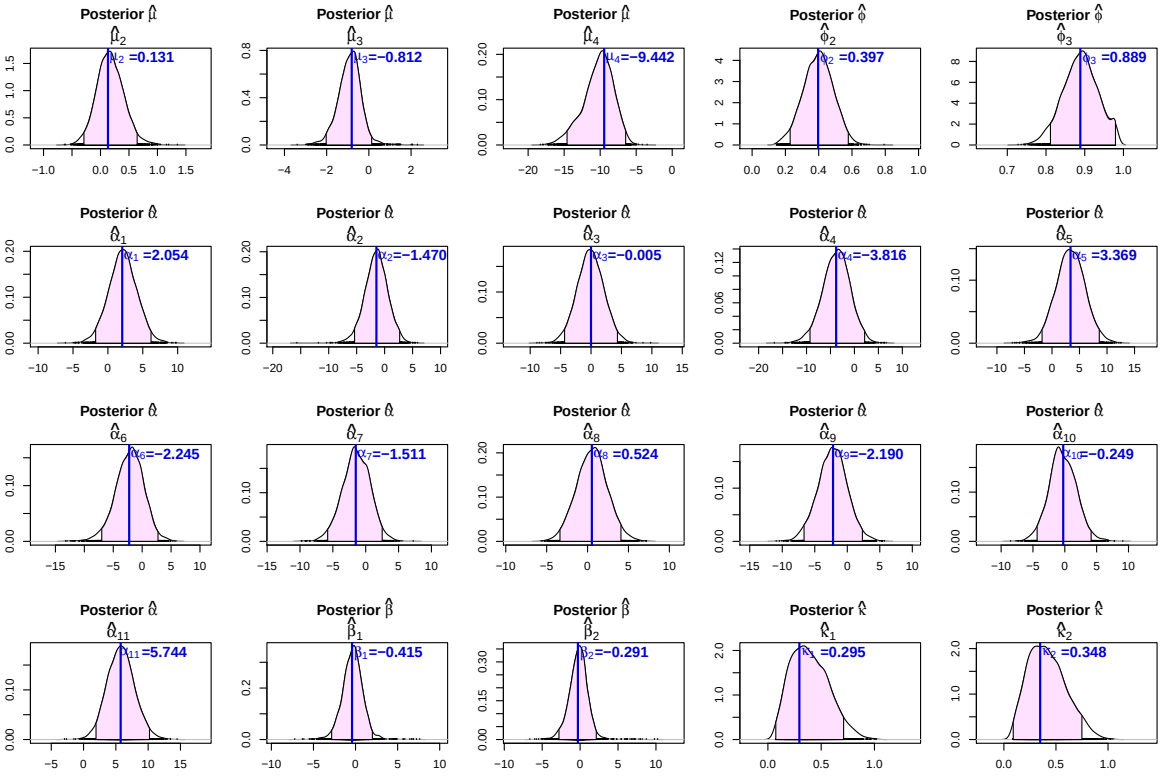

Figure 8.10: ***Spider data set:*** *Density plot depicting the marginal posterior distribution of the parameters for stereotype model including column clustering model* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_i} + \boldsymbol{\beta_c})$ *with* $C = 3$ *column (site) clusters. The blue vertical lines are the MLE values. 95% HPD credible intervals are shown with shading area.*

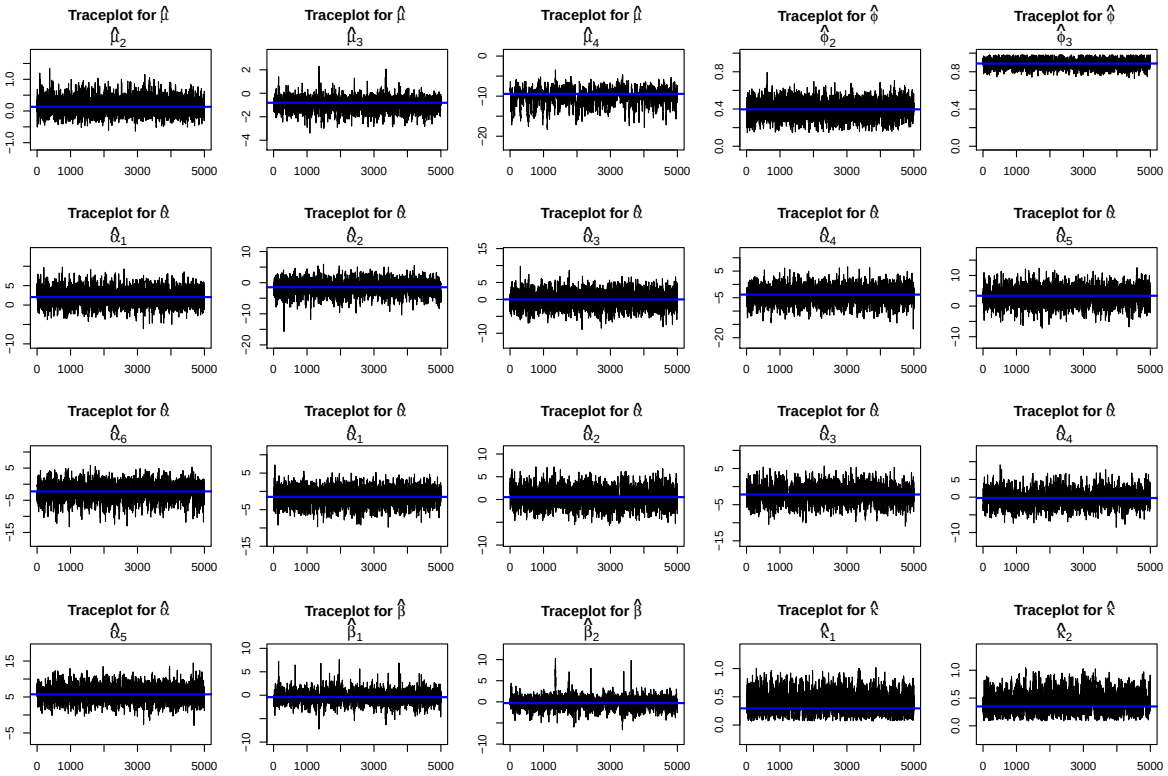

Figure 8.11: ***Spider data set:*** *Trace plot of the parameters for stereotype model including column clustering model* $\boldsymbol{\mu_k + \phi_k(\alpha_i + \beta_c)}$ *with* $C = 3$ *column (site) clusters. The blue horizontal lines are the MLE values.*

Chapter 9

Conclusions and Future Research Directions

We have developed a set of likelihood models based on the ordinal stereotype model and have introduced fuzzy clustering via finite mixtures in order to reduce the dimensionality of the problem and simplify its interpretation. In the research literature, there are numerous methodologies dealing with the clustering of data in fields where multivariate techniques are necessary. The advantage of our approach is its likelihood-based foundation because maximum likelihood theory provides estimators and model selection. In addition, the fitted spacing $\{\widehat{\phi}_k\}$ among ordinal categories of the response variable is dictated by the data and arises from the use of the ordinal stereotype model. This is an advantage over other ordinal-based models such as the proportional odds model and the continuation-ratio model, which do not provide a direct interpretation of the spacing between ordinal levels. More research in performance comparison with others equivalent methods is needed and may be a direction for future research.

We have described two procedures to fit the suite of developed models. The first procedure is based on deriving the maximum likelihood estimators of the parameters using the EM algorithm. The second procedure is based on a Bayesian inference approach which has been implemented in two different methodologies whether if we consider the number of components in the mixture as a known value or an unknown parameter. In the first methodology, the models were fitted using the MCMC techniques through the Metropolis-Hastings algorithm and the fitting was achieved using the reversible jump MCMC sampler in the second case.

We have demonstrated both procedures by means of three examples with real-life data. We have considered the case where responses in each column have the same number of ordinal response levels. This can be varied but may require a separate set of parameters $\{\mu_{jk}\}$ and $\{\phi_{jk}\}$ in the formulation of the models. All the programs in this thesis have been written in **R**. Computation can be slow in this stage, though we have substantially reduced the time required by calling compiled **C** code from **R**.

The reliability of our methodology was tested through a simulation study using the EM algorithm and the two Bayesian inference methodologies. We detected that there is an indication of multimodality of the likelihood surface. One way to deal with this is to implement a convergence strategy where several starting values are tested over the parameter vector in order to obtain the global maximum.

In a frequentist framework, we tested model comparison conducting a simulation study in which the results show that AIC, AIC_c and AIC_u are effective information criteria to score fitted models based on our likelihood-based finite mixture model approach for ordinal datasets. In particular, AIC consistently performs best among the tested information criteria to select the model with the correct number of clusters in a wide range of scenarios. In the research literature, there has so far been minimal research on model selection for finite mixture models with categorical data. Thus, this was an empirical simulation study and the conclusions are based on a set of information criteria in common use, none of which were developed for ordinal data. Additionally, we used the same information criterion measures for the biclustering case as in the one-dimensional case. We are aware that the asymptotic properties which apply to the one-dimensional case might not apply in the two-dimensional case. Thus, the information criteria affect the two clusterings differently. Because of this, development of a specific measure for model comparison with ordinal variables and for the two-dimensional case is required and should be achieved in future research.

In a Bayesian framework, we implemented methodologies to diagnose the convergence of the MCMC samplers. Moreover, the selection of the best model was achieved using DIC, in the fixed-dimensional case. In addition, we implemented a relabelling algorithm to deal with the label switching problem. This problem arises when MCMC samplers are applied to parameter estimation and clustering using mixture models. The use of RJMCMC jointly with a label switch-

ing procedure allows selection of the best model dimension while the sampler is estimating the model parameters. This is an advantage over fixed-dimensional MCMC samplers (e.g. Metropolis-Hastings algorithm) where the different models to test have to be sampled independently. Based on our experience, two drawbacks of the RJMCMC sampler are that it requires accuracy in the selection of suitable proposal distributions and the mixing might be slower than in fixed-dimensional MCMC samplers. This thesis was focused on the development of a RJMCMC algorithm to sample from one-dimensional mixture models. Thus, future developments in this area would include the development of an extra layer in our RJMCMC sampler allowing both jumps between different class families (i.e., between with and without interaction models from the same family) and jumps between one-dimensional and two-dimensional models (biclustering).

We have presented new data visualisation methods for depicting the results of our approach. Output from our models allows us to determine a new spacing of the ordinal categories, dictated by the data. Thus, these graphical tools used this more appropriate spacing to lead to more informative visualisation. In particular, we developed multidimensional scaling plots, ordination plots, mosaic plots including the new spacing, and level plots which depict the fuzzy probabilistic clustering. These graphical tools allow us to present the dimensional reduction results visually, to understand them more easily, and to lead to the identification of patterns in the data. The development of visualisation techniques focused on the biclustering case may warrant future research. A possible direction would be to develop mixture-based biplots as described in Pledger and Arnold (2014) to represent associations among rows, row groups, columns and column groups.

A strategy of categorising count data into ordinal data was carried out and a set of measures to compare different cluster structures were implemented. Count data sets may involve overdispersion from a set of species and underdispersion from another set which would require to fit different models (e.g. a negative binomial model for the overdispersed set and a binomial model for the underdispersed one). In our view, the main advantage of using our ordinal approach is that it allows for the inclusion of all of the different variance-mean ratio cases in the data in one methodology, without treating the data differently. Additionally, many count data sets have very high counts and very low counts. Categorising these counts into ordinal categories makes the actual counts less influential in the model fitting, giving broad categories which enable us to detect major overall

patterns. Two future research directions may be set in order to investigate the differences between recoded and original count data. Firstly, setting up an empirical comprehensive study through numerical experiments across a wide range of scenarios. Secondly, developing a measure to quantify the loss of information due to use of the ordinal categorisation instead of the original count data.

Part of the work presented in this thesis (Chapters 1-4) was published in Fernández et al. (2014a). Other more general future research directions to consider will be:

- Create an **R** package including the finite mixture models for ordinal data introduced in this thesis. This package will include routines to fit models based on the EM algorithm and Bayesian inference approaches and to depict the clustering results using the visualisation tools introduced in Chapter 5.

- Develop measures of goodness of fit for our ordinal stereotype approach based on the inspection of the discrepancy between the observed and the expected values under the model. The detection of patterns in the residuals may indicate a stronger association than is predicted by the fitted model.

- Compare results of our approach versus other similar methodologies such as k-means and the proportional odds model version of the cumulative logit. In order to do that, we will simulate a continuous latent variable from a logistic or normal distribution and chop it into categories to get the ordinal response. Then, we will run the different methodologies with this ordinal data and compare the observed results between the different approaches. This comparison will allow us to determine the scenarios where our methodology based on the stereotype model has better results than other comparable approaches.

- Extend our finite mixture approach to other ordinal models such as logit models, loglinear models, cumulative link models and association models that do not have loglinear structure but can describe ordinal association.

- Investigate models involving other correlation structures such as column dependence based on repeated measurements, i.e. longitudinal data. We

could then generalize the approach to the development of clustering methods of three-way data, i.e. biclustering individuals and questions, with repeated measures.

- Include row or column covariates such as environmental site factors in ecological data, or respondent characteristics in questionnaire data.
- Regarding the number of components in the mixture, we are interested in exploring approaches where the number of components is not fixed in advance. It might be possible to consider a Dirichlet process prior in a Bayesian nonparametric framework. This assumes that an infinite number of components exists, but most of them are never seen. The method would be created for including, adding or reducing components in the process.

Appendices

Appendix A

Model Formulation. Column Clustering and Biclustering

A.1 Response Probabilities in the Ordered Stereotype Model

In this appendix, we describe the relationship between models (2.3) and (2.6), which were formulated in Section 2.3.

From equation (2.6) with the predictor including the covariates and $\sum_{k=1}^{q} P\left[y_{ij} = k \mid \boldsymbol{x}\right] = 1$, we calculate

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

Using the identifiability constraints $\mu_1 = \phi_1 = 0$ it follows that

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

Therefore, equation (2.3) can be obtained from (2.6) just using the above expression (A.1) for $P\left[y_{ij} = 1 \mid \boldsymbol{x}\right]$.

A.2 EM Algorithm Formulae. Column clustering

In section 2.5, we described the model fitting procedure for the row clustering case. In this appendix, the fitting procedure is formulated for the case of column clustering.

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

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

E-step:

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

and

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

M-step:

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

and

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

conditional on the identifiability constraints on the parameters.

A.3 EM Algorithm Formulae. Biclustering

In Section 2.5, we described the model fitting procedure for the row clustering case. In this appendix, the fitting procedure is formulated for the case of biclustering.

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

E-step:

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

and

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

The E-step of the EM algorithm calls for the expected value of the complete data log-likelihood taking into account the fact that the only data unknown is $\{z_{ir}\}$

and $\{x_{jc}\}$ conditional on the observed data $\{y_{ij}\}$:

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

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

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

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

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

$$\begin{aligned}\widehat{Q}(\Omega \mid \Omega^{(t-1)}) = &\sum_{i=1}^{n}\sum_{r=1}^{R} \widehat{Z}_{ir} \log\left(\widehat{\pi}_r^{(t-1)}\right) + \sum_{j=1}^{m}\sum_{c=1}^{C} \widehat{X}_{jc} \log\left(\widehat{\kappa}_c^{(t-1)}\right) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q}\sum_{r=1}^{R}\sum_{c=1}^{C} \widehat{Z}_{ir}\widehat{X}_{jc} I(y_{ij}=k) \log\left(\widehat{\theta}_{rck}^{(t-1)}\right).\end{aligned} \tag{A.2}$$

M-step:

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

and

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

and

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

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

The variational approximation presents several drawbacks (see e.g. Keribin et al. (2012) for a discussion on this topic). In our work, we have not employed the variational approximation for the ultimate MLEs. Instead, we have used an alternative procedure with the aim of ensuring a solution avoiding approximation. Thus, the MLEs from the EM algorithm are used as starting points in order to numerically maximise the incomplete-data log-likelihood (2.14) (or (2.15)).

Appendix B

Model Comparison. Results

B.1 Parameter Configuration

Tables B.1 and B.2 summarise the parameter configuration for each scenario in the row clustering and biclustering cases.

B.2 Row Clustering and Biclustering Results

The full results for all the scenarios broken down by number of rows/columns and sample size is given from Table B.3 to Table B.7 for the row clustering case and from Table B.8 to Table B.12 for the biclustering case.

B.3 Questionnaire. Three Cultures

The full set of questions in (Anders and Batchelder, 2013) study is given in Table B.13. The data set with the responses of 83 respondents over 20 questions is given in Table B.14.

Table B.1: Parameter configuration for 5 tested scenarios in the row clustering case.

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

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

Table B.2: Parameter configuration for 5 tested scenarios in the biclustering case.

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

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

Table B.3: Model comparison simulation study results for 11 information criteria. Row Clustering ($\mu_k + \phi_k(\alpha_r + \beta_j)$). Scenario 1.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	7.9	37.7	9.1	20.3	66.4	57.9	62.8	62.7	44.3	66.0	66.1
		Fit	91.4	61.7	90.2	79.0	32.4	41.2	36.1	36.1	36.4	33.1	32.9
		Overfit	0.7	0.7	0.7	0.7	1.1	0.9	1.1	1.2	19.2	0.9	1.0
Sample size	$n = 50$	Underfit	18.7	54.7	18.7	43.3	66.0	61.3	65.3	56.7	42.7	65.3	65.3
		Fit	80.3	44.3	80.3	55.7	32.3	37.7	33.3	41.7	41.3	33.7	33.3
		Overfit	1.0	1.0	1.0	1.0	1.7	1.0	1.3	1.7	16.0	1.0	1.3
	$n = 100$	Underfit	3.7	34.7	5.0	10.3	66.7	58.7	66.0	64.7	46.7	66.0	66.3
		Fit	95.3	64.3	94.0	88.7	32.3	40.3	32.7	34.0	38.7	33.0	33.0
		Overfit	1.0	1.0	1.0	1.0	1.0	1.0	1.3	1.3	14.7	1.0	0.7
	$n = 500$	Underfit	1.3	23.7	3.7	7.3	66.7	53.7	57.0	66.7	43.7	66.7	66.7
		Fit	98.7	76.3	96.3	92.7	32.7	45.7	42.3	32.7	29.3	32.7	32.3
		Overfit	0.0	0.0	0.0	0.0	0.7	0.7	0.7	0.7	27.0	0.7	1.0
Number of row clusters	$R = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	98.0	98.0	98.0	96.7	97.3	96.7	96.3	61.7	97.3	97.0
		Overfit	2.0	2.0	2.0	2.0	3.3	2.7	3.3	3.7	38.3	2.7	3.0
	$R = 3$	Underfit	15.7	30.7	15.0	20.3	99.3	73.7	88.3	91.3	50.7	98.0	99.0
		Fit	84.3	69.3	85.0	79.7	0.7	26.3	11.7	8.7	35.3	2.0	1.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	0.0	0.0
	$R = 4$	Underfit	8.0	82.3	12.3	40.7	100.0	100.0	100.0	96.7	82.3	100.0	99.3
		Fit	92.0	17.7	87.7	59.3	0.0	0.0	0.0	3.3	12.3	0.0	0.7
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.3	0.0	0.0
$R = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	97.0	97.0	97.0	97.0	95.0	97.0	96.0	95.0	64.0	97.0	96.0
		Overfit	3.0	3.0	3.0	3.0	5.0	3.0	4.0	5.0	36.0	3.0	4.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	97.0	97.0	97.0	97.0	97.0	97.0	96.0	96.0	66.0	97.0	98.0
		Overfit	3.0	3.0	3.0	3.0	3.0	3.0	4.0	4.0	34.0	3.0	2.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	98.0	98.0	98.0	98.0	55.0	98.0	97.0
		Overfit	0.0	0.0	0.0	0.0	2.0	2.0	2.0	2.0	45.0	2.0	3.0
$R = 3$	$n = 50$	Underfit	44.0	70.0	42.0	58.0	98.0	84.0	96.0	78.0	38.0	96.0	98.0
		Fit	56.0	30.0	58.0	42.0	2.0	16.0	4.0	22.0	50.0	4.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.0	0.0	0.0
	$n = 100$	Underfit	3.0	22.0	3.0	3.0	100.0	76.0	98.0	96.0	46.0	98.0	99.0
		Fit	97.0	78.0	97.0	97.0	0.0	24.0	2.0	4.0	44.0	2.0	1.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	100.0	61.0	71.0	100.0	68.0	100.0	100.0
		Fit	100.0	100.0	100.0	100.0	0.0	39.0	29.0	0.0	12.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0
$R = 4$	$n = 50$	Underfit	12.0	94.0	14.0	72.0	100.0	100.0	100.0	92.0	90.0	100.0	98.0
		Fit	88.0	6.0	86.0	28.0	0.0	0.0	0.0	8.0	10.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$n = 100$	Underfit	8.0	82.0	12.0	28.0	100.0	100.0	100.0	98.0	94.0	100.0	100.0
		Fit	92.0	18.0	88.0	72.0	0.0	0.0	0.0	2.0	6.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$n = 500$	Underfit	4.0	71.0	11.0	22.0	100.0	100.0	100.0	100.0	63.0	100.0	100.0
		Fit	96.0	29.0	89.0	78.0	0.0	0.0	0.0	0.0	21.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.0	0.0	0.0

Notes: All the data is shown in percentage form(%).

Table B.4: Model comparison simulation study results for 11 information criteria. Row Clustering ($\mu_k + \phi_k(\alpha_r + \beta_j)$). Scenario 2.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	2.0	34.2	4.9	19.8	66.4	60.9	63.3	62.2	51.8	66.0	66.4
		Fit	97.6	65.6	94.8	80.0	33.6	39.1	36.7	37.8	36.0	34.0	32.9
		Overfit	0.4	0.2	0.3	0.2	0.0	0.0	0.0	0.0	12.2	0.0	0.7
Sample size	$n = 50$	Underfit	4.0	36.7	7.3	28.0	66.0	61.3	64.0	54.0	40.0	64.7	66.7
		Fit	96.0	63.3	92.7	72.0	34.0	38.7	36.0	46.0	44.0	35.3	31.3
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.0	0.0	2.0
	$n = 100$	Underfit	0.0	42.0	4.7	28.7	66.7	62.7	64.0	66.0	57.3	66.7	66.0
		Fit	99.3	58.0	94.7	71.3	33.3	37.3	36.0	34.0	30.0	33.3	34.0
		Overfit	0.7	0.0	0.7	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0
	$n = 500$	Underfit	2.0	24.0	2.7	2.7	66.7	58.7	62.0	66.7	58.0	66.7	66.7
		Fit	97.3	75.3	97.0	96.7	33.3	41.3	38.0	33.3	34.0	33.3	33.3
		Overfit	0.7	0.7	0.3	0.7	0.0	0.0	0.0	0.0	8.0	0.0	0.0
Number of row clusters	$R = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	74.7	100.0	98.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.3	0.0	2.0
	$R = 3$	Underfit	3.3	19.3	5.3	10.7	99.3	84.0	90.7	91.3	63.3	98.0	100.0
		Fit	95.3	80.0	93.7	88.7	0.7	16.0	9.3	8.7	26.0	2.0	0.0
		Overfit	1.3	0.7	1.0	0.7	0.0	0.0	0.0	0.0	10.7	0.0	0.0
	$R = 4$	Underfit	2.7	83.3	9.3	48.7	100.0	98.7	99.3	95.3	92.0	100.0	99.3
		Fit	97.3	16.7	90.7	51.3	0.0	1.3	0.7	4.7	7.3	0.0	0.7
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0
$R = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	68.0	100.0	94.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.0	0.0	6.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0	100.0	100.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	30.0	0.0	0.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	86.0	100.0	100.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	0.0	0.0
$R = 3$	$n = 50$	Underfit	10.0	26.0	16.0	18.0	98.0	86.0	92.0	76.0	38.0	94.0	100.0
		Fit	90.0	74.0	84.0	82.0	2.0	14.0	8.0	24.0	46.0	6.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.0	0.0	0.0
	$n = 100$	Underfit	0.0	32.0	0.0	14.0	100.0	88.0	92.0	98.0	76.0	100.0	100.0
		Fit	98.0	68.0	98.0	86.0	0.0	12.0	8.0	2.0	16.0	0.0	0.0
		Overfit	2.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	8.0	0.0	0.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	100.0	78.0	88.0	100.0	76.0	100.0	100.0
		Fit	98.0	98.0	99.0	98.0	0.0	22.0	12.0	0.0	16.0	0.0	0.0
		Overfit	2.0	2.0	1.0	2.0	0.0	0.0	0.0	0.0	8.0	0.0	0.0
$R = 4$	$n = 50$	Underfit	2.0	84.0	6.0	66.0	100.0	98.0	100.0	86.0	82.0	100.0	100.0
		Fit	98.0	16.0	94.0	34.0	0.0	2.0	0.0	14.0	18.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$n = 100$	Underfit	0.0	94.0	14.0	72.0	100.0	100.0	100.0	100.0	96.0	100.0	98.0
		Fit	100.0	6.0	86.0	28.0	0.0	0.0	0.0	0.0	4.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$n = 500$	Underfit	6.0	72.0	8.0	8.0	100.0	98.0	98.0	100.0	98.0	100.0	100.0
		Fit	94.0	28.0	92.0	92.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0

Notes: All the data is shown in percentage form(%).

Table B.5: Model comparison simulation study results for 11 information criteria. Row Clustering ($\mu_k + \phi_k(\alpha_r + \beta_j)$). Scenario 3.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	10.0	42.4	23.3	31.3	66.0	60.0	62.7	60.2	42.0	65.3	65.1
		Fit	88.0	56.7	74.7	66.7	31.1	40.0	37.3	34.9	38.7	30.2	31.1
		Overfit	2.0	0.9	2.0	2.0	2.9	0.0	0.0	4.9	19.3	4.4	3.8
Sample size	$n = 50$	Underfit	5.3	51.3	24.7	43.3	66.7	64.7	66.7	57.3	41.3	66.0	63.3
		Fit	94.0	48.0	74.7	56.0	31.3	35.3	33.3	37.3	36.7	30.0	27.3
		Overfit	0.7	0.7	0.7	0.7	2.0	0.0	0.0	5.3	22.0	4.0	9.3
	$n = 100$	Underfit	11.7	48.7	32.0	36.0	66.0	63.3	65.3	58.0	40.7	64.7	65.3
		Fit	88.3	51.3	68.0	64.0	32.0	36.7	34.7	38.7	38.7	32.0	32.7
		Overfit	0.0	0.0	0.0	0.0	2.0	0.0	0.0	3.3	20.7	3.3	2.0
	$n = 500$	Underfit	13.0	27.3	13.3	14.7	65.3	52.0	56.0	65.3	44.0	65.3	66.7
		Fit	81.7	70.7	81.3	80.0	30.0	48.0	44.0	28.7	40.7	28.7	33.3
		Overfit	5.3	2.0	5.3	5.3	4.7	0.0	0.0	6.0	15.3	6.0	0.0
Number of row clusters	$R = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	96.7	98.0	96.7	96.7	91.3	100.0	100.0	85.3	61.3	86.7	88.7
		Overfit	3.3	2.0	3.3	3.3	8.7	0.0	0.0	14.7	38.7	13.3	11.3
	$R = 3$	Underfit	0.0	42.7	7.3	25.3	98.0	81.3	88.7	82.7	34.7	96.0	96.7
		Fit	98.0	56.7	90.7	72.7	2.0	18.7	11.3	17.3	47.3	4.0	3.3
		Overfit	2.0	0.7	2.0	2.0	0.0	0.0	0.0	0.0	18.0	0.0	0.0
	$R = 4$	Underfit	30.0	84.7	62.7	68.7	100.0	98.7	99.3	98.0	91.3	100.0	98.7
		Fit	69.3	15.3	36.7	30.7	0.0	1.3	0.7	2.0	7.3	0.0	1.3
		Overfit	0.7	0.0	0.7	0.7	0.0	0.0	0.0	0.0	1.3	0.0	0.0
$R = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	98.0	98.0	98.0	94.0	100.0	100.0	84.0	56.0	88.0	72.0
		Overfit	2.0	2.0	2.0	2.0	6.0	0.0	0.0	16.0	44.0	12.0	28.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	94.0	100.0	100.0	90.0	56.0	90.0	94.0
		Overfit	0.0	0.0	0.0	0.0	6.0	0.0	0.0	10.0	44.0	10.0	6.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	92.0	96.0	92.0	92.0	86.0	100.0	100.0	82.0	72.0	82.0	100.0
		Overfit	8.0	4.0	8.0	8.0	14.0	0.0	0.0	18.0	28.0	18.0	0.0
$R = 3$	$n = 50$	Underfit	0.0	64.0	0.0	46.0	100.0	94.0	100.0	76.0	34.0	98.0	92.0
		Fit	100.0	36.0	100.0	54.0	0.0	6.0	0.0	24.0	46.0	2.0	8.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0
	$n = 100$	Underfit	0.0	54.0	22.0	30.0	98.0	90.0	96.0	76.0	30.0	94.0	98.0
		Fit	100.0	46.0	78.0	70.0	2.0	10.0	4.0	24.0	54.0	6.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.0	0.0	0.0
	$n = 500$	Underfit	0.0	10.0	0.0	0.0	96.0	60.0	70.0	96.0	40.0	96.0	100.0
		Fit	94.0	88.0	94.0	94.0	4.0	40.0	30.0	4.0	42.0	4.0	0.0
		Overfit	6.0	2.0	6.0	6.0	0.0	0.0	0.0	0.0	18.0	0.0	0.0
$R = 4$	$n = 50$	Underfit	16.0	90.0	74.0	84.0	100.0	100.0	100.0	96.0	90.0	100.0	98.0
		Fit	84.0	10.0	26.0	16.0	0.0	0.0	0.0	4.0	8.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
	$n = 100$	Underfit	35.0	92.0	74.0	78.0	100.0	100.0	100.0	98.0	92.0	100.0	98.0
		Fit	65.0	8.0	26.0	22.0	0.0	0.0	0.0	2.0	6.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
	$n = 500$	Underfit	39.0	72.0	40.0	44.0	100.0	96.0	98.0	100.0	92.0	100.0	100.0
		Fit	59.0	28.0	58.0	54.0	0.0	4.0	2.0	0.0	8.0	0.0	0.0
		Overfit	2.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Notes: All the data is shown in percentage form(%).

Table B.6: Model comparison simulation study results for 11 information criteria. Row Clustering ($\mu_k + \phi_k(\alpha_r + \beta_j)$). Scenario 4.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	5.8	43.1	7.6	10.7	66.7	60.2	62.9	61.3	45.3	66.2	65.6
		Fit	92.9	56.4	91.1	88.0	33.3	39.6	37.1	38.4	36.7	33.8	33.6
		Overfit	1.3	0.4	1.3	1.3	0.0	0.2	0.0	0.2	18.0	0.0	0.9
Sample size	$n = 50$	Underfit	14.7	49.3	17.3	22.7	66.7	64.0	65.3	53.3	37.3	65.3	64.0
		Fit	85.3	50.7	82.7	77.3	33.3	36.0	34.7	46.7	48.0	34.7	34.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.7	0.0	2.0
	$n = 100$	Underfit	0.0	44.7	2.7	6.7	66.7	60.0	62.7	64.0	42.0	66.7	66.0
		Fit	100.0	55.3	97.3	93.3	33.3	40.0	37.3	35.3	38.7	33.3	33.3
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	19.3	0.0	0.7
	$n = 500$	Underfit	2.7	35.3	2.7	2.7	66.7	56.7	60.7	66.7	56.7	66.7	66.7
		Fit	93.3	63.3	93.3	93.3	33.3	42.7	39.3	33.3	23.3	33.3	33.3
		Overfit	4.0	1.3	4.0	4.0	0.0	0.7	0.0	0.0	20.0	0.0	0.0
Number of row clusters	$R = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.7	99.3	98.7	98.7	100.0	99.3	100.0	99.3	60.0	100.0	98.7
		Overfit	1.3	0.7	1.3	1.3	0.0	0.7	0.0	0.7	40.0	0.0	1.3
	$R = 3$	Underfit	14.7	40.7	18.0	24.7	100.0	81.3	89.3	90.7	50.7	98.7	98.0
		Fit	82.7	58.7	79.3	72.7	0.0	18.7	10.7	9.3	36.0	1.3	0.7
		Overfit	2.7	0.7	2.7	2.7	0.0	0.0	0.0	0.0	13.3	0.0	1.3
	$R = 4$	Underfit	2.7	88.7	4.7	7.3	100.0	99.3	99.3	93.3	85.3	100.0	98.7
		Fit	97.3	11.3	95.3	92.7	0.0	0.7	0.7	6.7	14.0	0.0	1.3
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0
$R = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	72.0	100.0	96.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	0.0	4.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0	58.0	100.0	100.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	42.0	0.0	0.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	96.0	98.0	96.0	96.0	100.0	98.0	100.0	100.0	50.0	100.0	100.0
		Overfit	4.0	2.0	4.0	4.0	0.0	2.0	0.0	0.0	50.0	0.0	0.0
$R = 3$	$n = 50$	Underfit	38.0	58.0	40.0	48.0	100.0	92.0	96.0	74.0	36.0	96.0	96.0
		Fit	62.0	42.0	60.0	52.0	0.0	8.0	4.0	26.0	50.0	4.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	0.0	2.0
	$n = 100$	Underfit	0.0	42.0	8.0	20.0	100.0	82.0	90.0	98.0	46.0	100.0	98.0
		Fit	100.0	58.0	92.0	80.0	0.0	18.0	10.0	2.0	38.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.0	0.0	2.0
	$n = 500$	Underfit	6.0	22.0	6.0	6.0	100.0	70.0	82.0	100.0	70.0	100.0	100.0
		Fit	86.0	76.0	86.0	86.0	0.0	30.0	18.0	0.0	20.0	0.0	0.0
		Overfit	8.0	2.0	8.0	8.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0
$R = 4$	$n = 50$	Underfit	6.0	90.0	12.0	20.0	100.0	100.0	100.0	86.0	76.0	100.0	96.0
		Fit	94.0	10.0	88.0	80.0	0.0	0.0	0.0	14.0	22.0	0.0	4.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
	$n = 100$	Underfit	0.0	92.0	0.0	0.0	100.0	98.0	98.0	94.0	80.0	100.0	100.0
		Fit	100.0	8.0	100.0	100.0	0.0	2.0	2.0	6.0	20.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$n = 500$	Underfit	2.0	84.0	2.0	2.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		Fit	98.0	16.0	98.0	98.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Notes: All the data is shown in percentage form(%).

Table B.7: Model comparison simulation study results for 11 information criteria. Row Clustering ($\mu_k + \phi_k(\alpha_r + \beta_j)$). Scenario 5.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	0.0	0.4	0.0	0.0	55.6	40.9	42.0	59.8	56.2	65.8	66.2
		Fit	99.1	98.2	98.2	98.2	33.3	58.7	57.6	40.0	33.8	34.2	32.9
		Overfit	0.9	1.3	1.8	1.8	0.0	0.4	0.4	0.2	10.0	0.0	0.9
Sample size	$n = 50$	Underfit	0.0	0.0	0.0	0.0	33.3	48.0	46.7	51.3	50.0	64.7	66.7
		Fit	98.7	100.0	99.3	99.3	33.3	52.0	53.3	48.0	38.0	35.3	32.0
		Overfit	1.3	0.0	0.7	0.7	0.0	0.0	0.0	0.7	12.0	0.0	1.3
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	66.7	42.0	46.7	62.0	57.3	66.0	66.0
		Fit	100.0	100.0	100.0	100.0	33.3	58.0	53.3	38.0	32.0	34.0	32.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.7	0.0	1.3
	$n = 500$	Underfit	0.0	1.3	0.0	0.0	66.7	32.7	32.7	66.0	61.3	66.7	66.0
		Fit	98.7	94.7	95.3	95.3	33.3	66.0	66.0	34.0	31.3	33.3	34.0
		Overfit	1.3	4.0	4.7	4.7	0.0	1.3	1.3	0.0	7.3	0.0	0.0
Number of row clusters	$R = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	99.3	100.0	100.0	100.0	100.0	100.0	100.0	99.3	72.7	100.0	97.3
		Overfit	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.7	27.3	0.0	2.7
	$R = 3$	Underfit	0.0	0.0	0.0	0.0	100.0	26.0	28.0	89.3	73.3	98.7	100.0
		Fit	98.0	96.0	94.7	94.7	0.0	72.7	70.7	10.7	24.0	1.3	0.0
		Overfit	2.0	4.0	5.3	5.3	0.0	1.3	1.3	0.0	2.7	0.0	0.0
	$R = 4$	Underfit	0.0	1.3	0.0	0.0	66.7	96.7	98.0	90.0	95.3	98.7	98.7
		Fit	100.0	98.7	100.0	100.0	0.0	3.3	2.0	10.0	4.7	1.3	1.3
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$R = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0	70.0	100.0	96.0
		Overfit	2.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	30.0	0.0	4.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0	100.0	96.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	30.0	0.0	4.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	78.0	100.0	100.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.0	0.0	0.0
$R = 3$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	100.0	48.0	44.0	72.0	56.0	98.0	100.0
		Fit	98.0	100.0	98.0	98.0	0.0	52.0	56.0	28.0	38.0	2.0	0.0
		Overfit	2.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	100.0	30.0	40.0	96.0	78.0	98.0	100.0
		Fit	100.0	100.0	100.0	100.0	0.0	70.0	60.0	4.0	20.0	2.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100.0	86.0	100.0	100.0
		Fit	96.0	88.0	86.0	86.0	0.0	96.0	96.0	0.0	14.0	0.0	0.0
		Overfit	4.0	12.0	14.0	14.0	0.0	4.0	4.0	0.0	0.0	0.0	0.0
$R = 4$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	100.0	96.0	96.0	82.0	94.0	96.0	100.0
		Fit	100.0	100.0	100.0	100.0	0.0	4.0	4.0	18.0	6.0	4.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	100.0	96.0	100.0	90.0	94.0	100.0	98.0
		Fit	100.0	100.0	100.0	100.0	0.0	4.0	0.0	10.0	6.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$n = 500$	Underfit	0.0	4.0	0.0	0.0	100.0	98.0	98.0	98.0	98.0	100.0	98.0
		Fit	100.0	96.0	100.0	100.0	0.0	2.0	2.0	2.0	2.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Notes: All the data is shown in percentage form(%).

Table B.8: Model comparison simulation study results for 11 information criteria. Biclustering ($\mu_k + \phi_k(\alpha_r + \beta_c)$). Scenario 1.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	9.5	23.5	9.7	14.0	64.2	65.3	68.8	50.0	15.2	58.2	61.7
		Fit	89.2	75.8	89.2	84.8	28.5	34.5	31.0	29.5	42.5	29.2	31.3
		Overfit	1.3	0.7	1.2	1.2	7.3	0.2	0.2	20.5	42.3	12.7	7.0
Sample size	$n = 50$	Underfit	2.0	13.3	2.0	2.0	57.3	54.0	57.3	26.7	6.0	42.7	44.0
		Fit	96.0	85.3	96.7	96.7	38.7	45.3	42.0	52.0	61.3	45.3	42.0
		Overfit	2.0	1.3	1.3	1.3	4.0	0.7	0.7	21.3	32.7	12.0	14.0
	$n = 100$	Underfit	5.3	8.0	5.3	6.0	51.3	55.3	60.7	37.3	8.7	46.0	51.3
		Fit	93.3	91.3	93.3	92.7	38.7	44.7	39.3	38.0	32.7	38.7	40.7
		Overfit	1.3	0.7	1.3	1.3	10.0	0.0	0.0	24.7	58.7	15.3	8.0
	$n = 500$	Underfit	1.3	3.3	1.3	1.3	48.0	52.0	57.3	40.0	13.3	44.0	54.0
		Fit	96.7	96.0	96.7	96.7	36.7	48.0	42.7	24.7	27.3	32.7	42.0
		Overfit	2.0	0.7	2.0	2.0	15.3	0.0	0.0	35.3	59.3	23.3	4.0
Number of row clusters	$R = 2$ $C = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	99.3	98.0	98.0	86.7	100.0	100.0	67.3	30.7	78.7	96.0
		Overfit	2.0	0.7	2.0	2.0	13.3	0.0	0.0	32.7	69.3	21.3	4.0
	$R = 2$ $C = 3$	Underfit	4.7	10.0	4.7	5.3	57.3	76.0	84.0	17.3	14.7	35.3	55.3
		Fit	93.3	88.7	93.3	92.7	26.7	24.0	16.0	36.0	40.7	36.0	25.3
		Overfit	2.0	1.3	2.0	2.0	16.0	0.0	0.0	46.7	44.7	28.7	19.3
	$R = 3$ $C = 2$	Underfit	4.0	14.7	4.0	4.0	99.3	85.3	91.3	86.7	13.3	97.3	94.0
		Fit	94.7	84.7	95.3	95.3	0.7	14.0	8.0	11.3	50.0	2.0	3.3
		Overfit	1.3	0.7	0.7	0.7	0.0	0.7	0.7	2.0	36.7	0.7	2.7
	$R = 3$ $C = 3$	Underfit	29.3	69.3	30.0	46.7	100.0	100.0	100.0	96.0	32.7	100.0	97.3
		Fit	70.7	30.7	70.0	53.3	0.0	0.0	0.0	3.3	48.7	0.0	0.7
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	18.7	0.0	2.0
$R = 2$ $C = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	98.0	98.0	98.0	92.0	100.0	100.0	78.0	86.0	84.0	94.0
		Overfit	2.0	2.0	2.0	2.0	8.0	0.0	0.0	22.0	14.0	16.0	6.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	100.0	98.0	98.0	88.0	100.0	100.0	78.0	2.0	84.0	96.0
		Overfit	2.0	0.0	2.0	2.0	12.0	0.0	0.0	22.0	98.0	16.0	4.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	100.0	98.0	98.0	80.0	100.0	100.0	46.0	4.0	68.0	98.0
		Overfit	2.0	0.0	2.0	2.0	20.0	0.0	0.0	54.0	96.0	32.0	2.0
$R = 2$ $C = 3$	$n = 50$	Underfit	4.0	8.0	4.0	4.0	74.0	82.0	88.0	10.0	16.0	36.0	46.0
		Fit	96.0	92.0	96.0	96.0	22.0	18.0	12.0	54.0	50.0	46.0	24.0
		Overfit	0.0	0.0	0.0	0.0	4.0	0.0	0.0	36.0	34.0	18.0	30.0
	$n = 100$	Underfit	6.0	14.0	6.0	8.0	54.0	76.0	86.0	20.0	22.0	38.0	58.0
		Fit	92.0	84.0	92.0	90.0	28.0	24.0	14.0	28.0	24.0	32.0	24.0
		Overfit	2.0	2.0	2.0	2.0	18.0	0.0	0.0	52.0	54.0	30.0	18.0
	$n = 500$	Underfit	4.0	8.0	4.0	4.0	44.0	70.0	78.0	22.0	6.0	32.0	62.0
		Fit	92.0	90.0	92.0	92.0	30.0	30.0	22.0	26.0	48.0	30.0	28.0
		Overfit	4.0	2.0	4.0	4.0	26.0	0.0	0.0	52.0	46.0	38.0	10.0
$R = 3$ $C = 2$	$n = 50$	Underfit	2.0	32.0	2.0	2.0	98.0	80.0	84.0	70.0	2.0	92.0	86.0
		Fit	94.0	66.0	96.0	96.0	2.0	18.0	14.0	24.0	48.0	6.0	8.0
		Overfit	4.0	2.0	2.0	2.0	0.0	2.0	2.0	6.0	50.0	2.0	6.0
	$n = 100$	Underfit	10.0	10.0	10.0	10.0	100.0	90.0	96.0	92.0	4.0	100.0	96.0
		Fit	90.0	90.0	90.0	90.0	0.0	10.0	4.0	8.0	72.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.0	0.0	2.0
	$n = 500$	Underfit	0.0	2.0	0.0	0.0	100.0	86.0	94.0	98.0	34.0	100.0	100.0
		Fit	100.0	98.0	100.0	100.0	0.0	14.0	6.0	2.0	30.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	36.0	0.0	0.0
$R = 3$ $C = 3$	$n = 50$	Underfit	78.0	92.0	80.0	82.0	100.0	100.0	100.0	94.0	0.0	100.0	96.0
		Fit	22.0	8.0	20.0	18.0	0.0	0.0	0.0	4.0	98.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	2.0	0.0	4.0
	$n = 100$	Underfit	4.0	68.0	4.0	52.0	100.0	100.0	100.0	98.0	44.0	100.0	96.0
		Fit	96.0	32.0	96.0	48.0	0.0	0.0	0.0	2.0	32.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.0	0.0	2.0
	$n = 500$	Underfit	6.0	48.0	6.0	6.0	100.0	100.0	100.0	96.0	54.0	100.0	100.0
		Fit	94.0	52.0	94.0	94.0	0.0	0.0	0.0	4.0	16.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	30.0	0.0	0.0

Notes: All the data is shown in percentage form(%).

Table B.9: Model comparison simulation study results for 11 information criteria. Biclustering ($\mu_k + \phi_k(\alpha_r + \beta_c)$). Scenario 2.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	16.5	33.7	17.3	18.2	61.8	64.8	68.2	52.3	15.2	57.7	58.8
		Fit	82.3	65.5	81.5	80.7	31.8	35.2	31.8	31.2	35.7	32.0	32.5
		Overfit	1.2	0.8	1.2	1.2	6.3	0.0	0.0	16.5	49.2	10.3	8.7
Sample size	$n = 50$	Underfit	4.0	16.0	4.0	4.7	56.0	56.7	62.7	34.0	7.3	45.3	44.0
		Fit	95.3	83.3	95.3	94.7	40.0	43.3	37.3	46.7	32.7	44.0	43.3
		Overfit	0.7	0.7	0.7	0.7	4.0	0.0	0.0	19.3	60.0	10.7	12.7
	$n = 100$	Underfit	2.0	14.7	2.0	2.0	47.3	54.7	58.0	41.3	3.3	43.3	46.0
		Fit	97.3	84.7	97.3	97.3	46.0	45.3	42.0	38.0	44.7	42.7	42.7
		Overfit	0.7	0.7	0.7	0.7	6.7	0.0	0.0	20.7	52.0	14.0	11.3
	$n = 500$	Underfit	0.7	18.7	0.7	0.7	44.0	49.3	52.7	38.7	12.7	42.0	47.3
		Fit	96.7	79.3	96.7	96.7	41.3	50.7	47.3	35.3	22.0	41.3	43.3
		Overfit	2.7	2.0	2.7	2.7	14.7	0.0	0.0	26.0	65.3	16.7	9.3
Number of row clusters	$R = 2$ $C = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.7	98.7	98.7	98.7	96.0	100.0	100.0	78.7	8.0	90.7	96.7
		Overfit	1.3	1.3	1.3	1.3	4.0	0.0	0.0	21.3	92.0	9.3	3.3
	$R = 2$ $C = 3$	Underfit	2.7	8.0	2.7	2.7	47.3	70.7	78.7	19.3	10.7	31.3	41.3
		Fit	94.7	90.0	94.7	94.7	31.3	29.3	21.3	36.0	46.0	36.7	31.3
		Overfit	2.7	2.0	2.7	2.7	21.3	0.0	0.0	44.7	43.3	32.0	27.3
	$R = 3$ $C = 2$	Underfit	4.0	41.3	4.0	4.7	100.0	90.0	94.7	94.7	12.7	99.3	96.0
		Fit	96.0	58.7	96.0	95.3	0.0	10.0	5.3	5.3	45.3	0.7	1.3
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	42.0	0.0	2.7
	$R = 3$ $C = 3$	Underfit	59.3	85.3	62.7	65.3	100.0	98.7	99.3	95.3	37.3	100.0	98.0
		Fit	40.0	14.7	36.7	34.0	0.0	1.3	0.7	4.7	43.3	0.0	0.7
		Overfit	0.7	0.0	0.7	0.7	0.0	0.0	0.0	0.0	19.3	0.0	1.3
$R = 2$ $C = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	98.0	98.0	98.0	98.0	100.0	100.0	82.0	18.0	94.0	98.0
		Overfit	2.0	2.0	2.0	2.0	2.0	0.0	0.0	18.0	82.0	6.0	2.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	98.0	98.0	98.0	96.0	100.0	100.0	72.0	6.0	84.0	92.0
		Overfit	2.0	2.0	2.0	2.0	4.0	0.0	0.0	28.0	94.0	16.0	8.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	94.0	100.0	100.0	82.0	0.0	94.0	100.0
		Overfit	0.0	0.0	0.0	0.0	6.0	0.0	0.0	18.0	100.0	6.0	0.0
$R = 2$ $C = 3$	$n = 50$	Underfit	4.0	14.0	4.0	4.0	68.0	82.0	94.0	14.0	10.0	38.0	42.0
		Fit	96.0	86.0	96.0	96.0	22.0	18.0	6.0	46.0	50.0	36.0	28.0
		Overfit	0.0	0.0	0.0	0.0	10.0	0.0	0.0	40.0	40.0	26.0	30.0
	$n = 100$	Underfit	4.0	6.0	4.0	4.0	42.0	74.0	78.0	26.0	8.0	30.0	40.0
		Fit	96.0	94.0	96.0	96.0	42.0	26.0	22.0	40.0	62.0	44.0	36.0
		Overfit	0.0	0.0	0.0	0.0	16.0	0.0	0.0	34.0	30.0	26.0	24.0
	$n = 500$	Underfit	0.0	4.0	0.0	0.0	32.0	56.0	64.0	18.0	14.0	26.0	42.0
		Fit	92.0	90.0	92.0	92.0	30.0	44.0	36.0	22.0	26.0	30.0	30.0
		Overfit	8.0	6.0	8.0	8.0	38.0	0.0	0.0	60.0	60.0	44.0	28.0
$R = 3$ $C = 2$	$n = 50$	Underfit	8.0	34.0	8.0	10.0	100.0	88.0	94.0	88.0	12.0	98.0	90.0
		Fit	92.0	66.0	92.0	90.0	0.0	12.0	6.0	12.0	30.0	2.0	4.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	58.0	0.0	6.0
	$n = 100$	Underfit	2.0	38.0	2.0	2.0	100.0	90.0	96.0	98.0	2.0	100.0	98.0
		Fit	98.0	62.0	98.0	98.0	0.0	10.0	4.0	2.0	66.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.0	0.0	2.0
	$n = 500$	Underfit	2.0	52.0	2.0	2.0	100.0	92.0	94.0	98.0	24.0	100.0	100.0
		Fit	98.0	48.0	98.0	98.0	0.0	8.0	6.0	2.0	40.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	36.0	0.0	0.0
$R = 3$ $C = 3$	$n = 50$	Underfit	68.0	94.0	76.0	78.0	100.0	98.0	100.0	92.0	58.0	100.0	96.0
		Fit	32.0	6.0	24.0	22.0	0.0	2.0	0.0	8.0	14.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	0.0	2.0
	$n = 100$	Underfit	64.0	86.0	66.0	72.0	100.0	98.0	98.0	96.0	52.0	100.0	100.0
		Fit	36.0	14.0	34.0	28.0	0.0	2.0	2.0	4.0	24.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.0	0.0	0.0
	$n = 500$	Underfit	46.0	76.0	46.0	46.0	100.0	100.0	100.0	98.0	2.0	100.0	98.0
		Fit	52.0	24.0	52.0	52.0	0.0	0.0	0.0	2.0	92.0	0.0	0.0
		Overfit	2.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0	6.0	0.0	2.0

Notes: All the data is shown in percentage form(%).

Table B.10: Model comparison simulation study results for 11 information criteria. Biclustering ($\mu_k + \phi_k(\alpha_r + \beta_c)$). Scenario 3.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	16.7	33.5	17.3	18.5	63.2	66.3	69.2	49.8	15.0	57.0	60.5
		Fit	80.5	64.7	80.0	79.3	30.3	33.5	30.7	33.8	40.8	31.2	30.5
		Overfit	2.8	1.8	2.7	2.2	6.5	0.2	0.2	16.3	44.2	11.8	9.0
Sample size	$n = 50$	Underfit	4.7	18.0	4.7	6.0	58.0	57.3	62.0	27.3	1.3	44.0	44.7
		Fit	92.0	80.7	92.7	92.0	37.3	42.0	37.3	50.7	56.0	43.3	37.3
		Overfit	3.3	1.3	2.7	2.0	4.7	0.7	0.7	22.0	42.7	12.7	18.0
	$n = 100$	Underfit	8.0	28.0	8.7	9.3	49.3	60.0	62.7	38.7	5.3	44.0	52.0
		Fit	90.0	70.0	89.3	88.7	41.3	40.0	37.3	42.0	44.0	42.0	41.3
		Overfit	2.0	2.0	2.0	2.0	9.3	0.0	0.0	19.3	50.7	14.0	6.7
	$n = 500$	Underfit	2.7	6.0	2.7	2.7	45.3	48.7	52.0	38.0	12.0	40.0	49.3
		Fit	92.0	90.0	92.0	92.7	42.7	51.3	48.0	38.0	24.7	39.3	43.3
		Overfit	5.3	4.0	5.3	4.7	12.0	0.0	0.0	24.0	63.3	20.7	7.3
Number of row clusters	$R = 2$ $C = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	97.3	99.3	98.0	98.0	92.0	100.0	100.0	82.7	5.3	87.3	93.3
		Overfit	2.7	0.7	2.0	2.0	8.0	0.0	0.0	17.3	94.7	12.7	6.7
	$R = 2$ $C = 3$	Underfit	9.3	31.3	9.3	11.3	54.7	78.7	82.7	15.3	10.7	31.3	48.0
		Fit	85.3	64.0	85.3	84.0	28.0	21.3	17.3	38.7	61.3	34.7	28.7
		Overfit	5.3	4.7	5.3	4.7	17.3	0.0	0.0	46.0	28.0	34.0	23.3
	$R = 3$ $C = 2$	Underfit	6.0	20.7	6.7	6.7	98.0	87.3	94.0	88.7	8.0	96.7	98.0
		Fit	91.3	77.3	90.7	91.3	1.3	12.0	5.3	9.3	58.0	2.7	0.0
		Overfit	2.7	2.0	2.7	2.0	0.7	0.7	0.7	2.0	34.0	0.7	2.0
	$R = 3$ $C = 3$	Underfit	51.3	82.0	53.3	56.0	100.0	99.3	100.0	95.3	41.3	100.0	96.0
		Fit	48.0	18.0	46.0	44.0	0.0	0.7	0.0	4.7	38.7	0.0	0.0
		Overfit	0.7	0.0	0.7	0.0	0.0	0.0	0.0	0.0	20.0	0.0	4.0
$R = 2$ $C = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	96.0	100.0	98.0	98.0	92.0	100.0	100.0	86.0	2.0	90.0	86.0
		Overfit	4.0	0.0	2.0	2.0	8.0	0.0	0.0	14.0	98.0	10.0	14.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	100.0	100.0	100.0	100.0	92.0	100.0	100.0	84.0	8.0	88.0	98.0
		Overfit	0.0	0.0	0.0	0.0	8.0	0.0	0.0	16.0	92.0	12.0	2.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	96.0	98.0	96.0	96.0	92.0	100.0	100.0	78.0	6.0	84.0	96.0
		Overfit	4.0	2.0	4.0	4.0	8.0	0.0	0.0	22.0	94.0	16.0	4.0
$R = 2$ $C = 3$	$n = 50$	Underfit	8.0	48.0	8.0	12.0	76.0	88.0	92.0	10.0	2.0	38.0	40.0
		Fit	92.0	52.0	92.0	88.0	20.0	12.0	8.0	42.0	82.0	36.0	26.0
		Overfit	0.0	0.0	0.0	0.0	4.0	0.0	0.0	48.0	16.0	26.0	34.0
	$n = 100$	Underfit	18.0	36.0	18.0	20.0	52.0	92.0	94.0	22.0	4.0	36.0	56.0
		Fit	76.0	58.0	76.0	74.0	28.0	8.0	6.0	38.0	80.0	34.0	26.0
		Overfit	6.0	6.0	6.0	6.0	20.0	0.0	0.0	40.0	16.0	30.0	18.0
	$n = 500$	Underfit	2.0	10.0	2.0	2.0	36.0	56.0	62.0	14.0	26.0	20.0	48.0
		Fit	88.0	82.0	88.0	90.0	36.0	44.0	38.0	36.0	22.0	34.0	34.0
		Overfit	10.0	8.0	10.0	8.0	28.0	0.0	0.0	50.0	52.0	46.0	18.0
$R = 3$ $C = 2$	$n = 50$	Underfit	6.0	6.0	6.0	6.0	98.0	84.0	94.0	72.0	2.0	94.0	94.0
		Fit	88.0	90.0	88.0	90.0	0.0	14.0	4.0	24.0	84.0	4.0	0.0
		Overfit	6.0	4.0	6.0	4.0	2.0	2.0	2.0	4.0	14.0	2.0	6.0
	$n = 100$	Underfit	6.0	48.0	8.0	8.0	96.0	88.0	94.0	94.0	12.0	96.0	100.0
		Fit	94.0	52.0	92.0	92.0	4.0	12.0	6.0	4.0	44.0	4.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	44.0	0.0	0.0
	$n = 500$	Underfit	6.0	8.0	6.0	6.0	100.0	90.0	94.0	100.0	10.0	100.0	100.0
		Fit	92.0	90.0	92.0	92.0	0.0	10.0	6.0	0.0	46.0	0.0	0.0
		Overfit	2.0	2.0	2.0	2.0	0.0	0.0	0.0	0.0	44.0	0.0	0.0
$R = 3$ $C = 3$	$n = 50$	Underfit	68.0	90.0	72.0	78.0	100.0	100.0	100.0	90.0	64.0	100.0	90.0
		Fit	30.0	10.0	26.0	22.0	0.0	0.0	0.0	10.0	18.0	0.0	0.0
		Overfit	2.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	18.0	0.0	10.0
	$n = 100$	Underfit	60.0	86.0	62.0	64.0	100.0	100.0	100.0	98.0	6.0	100.0	100.0
		Fit	40.0	14.0	38.0	36.0	0.0	0.0	0.0	2.0	92.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
	$n = 500$	Underfit	26.0	70.0	26.0	26.0	100.0	98.0	100.0	98.0	54.0	100.0	98.0
		Fit	74.0	30.0	74.0	74.0	0.0	2.0	0.0	2.0	6.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	40.0	0.0	2.0

Notes: All the data is shown in percentage form(%).

Table B.11: Model comparison simulation study results for 11 information criteria. Biclustering ($\mu_k + \phi_k(\alpha_r + \beta_c)$). Scenario 4.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	13.3	28.3	14.5	15.8	64.3	67.5	70.2	50.2	13.2	59.3	64.2
		Fit	85.5	70.8	84.5	83.3	28.2	32.3	29.7	30.5	35.0	27.3	29.2
		Overfit	1.2	0.8	1.0	0.8	7.5	0.2	0.2	19.3	51.8	13.3	6.7
Sample size	$n = 50$	Underfit	2.7	10.0	3.3	4.7	61.3	62.0	64.0	30.7	6.0	54.0	54.0
		Fit	96.7	90.0	96.0	95.3	34.0	38.0	36.0	48.7	47.3	35.3	34.0
		Overfit	0.7	0.0	0.7	0.0	4.7	0.0	0.0	20.7	46.7	10.7	12.0
	$n = 100$	Underfit	1.3	20.0	1.3	2.0	51.3	59.3	62.7	36.7	8.0	42.0	54.7
		Fit	96.7	78.0	96.7	96.0	38.7	40.0	36.7	34.0	23.3	38.0	36.7
		Overfit	2.0	2.0	2.0	2.0	10.0	0.7	0.7	29.3	68.7	20.0	8.7
	$n = 500$	Underfit	2.0	7.3	2.0	2.0	44.7	48.7	54.0	38.7	8.0	41.3	49.3
		Fit	96.7	91.3	96.7	96.7	40.0	51.3	46.0	34.0	28.0	36.0	45.3
		Overfit	1.3	1.3	1.3	1.3	15.3	0.0	0.0	27.3	64.0	22.7	5.3
Number of row clusters	$R = 2$ $C = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	97.3	98.0	97.3	98.0	89.3	99.3	99.3	66.7	4.7	80.0	90.7
		Overfit	2.7	2.0	2.7	2.0	10.7	0.7	0.7	33.3	95.3	20.0	9.3
	$R = 2$ $C = 3$	Underfit	3.3	14.0	3.3	5.3	58.0	81.3	87.3	21.3	6.7	39.3	62.0
		Fit	95.3	84.7	95.3	93.3	22.7	18.7	12.7	35.3	60.0	27.3	22.0
		Overfit	1.3	1.3	1.3	1.3	19.3	0.0	0.0	43.3	33.3	33.3	16.0
	$R = 3$ $C = 2$	Underfit	2.7	23.3	3.3	3.3	99.3	88.7	93.3	84.7	15.3	98.0	96.0
		Fit	97.3	76.7	96.7	96.7	0.7	11.3	6.7	14.7	34.0	2.0	3.3
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	50.7	0.0	0.7
	$R = 3$ $C = 3$	Underfit	47.3	76.0	51.3	54.7	100.0	100.0	100.0	94.7	30.7	100.0	98.7
		Fit	52.0	24.0	48.7	45.3	0.0	0.0	0.0	5.3	41.3	0.0	0.7
		Overfit	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	0.0	0.7
$R = 2$ $C = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	100.0	98.0	100.0	92.0	100.0	100.0	74.0	8.0	86.0	84.0
		Overfit	2.0	0.0	2.0	0.0	8.0	0.0	0.0	26.0	92.0	14.0	16.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	96.0	96.0	96.0	96.0	88.0	98.0	98.0	58.0	2.0	80.0	88.0
		Overfit	4.0	4.0	4.0	4.0	12.0	2.0	2.0	42.0	98.0	20.0	12.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	98.0	98.0	98.0	88.0	100.0	100.0	68.0	4.0	74.0	100.0
		Overfit	2.0	2.0	2.0	2.0	12.0	0.0	0.0	32.0	96.0	26.0	0.0
$R = 2$ $C = 3$	$n = 50$	Underfit	4.0	24.0	4.0	8.0	84.0	100.0	100.0	28.0	2.0	66.0	70.0
		Fit	96.0	76.0	96.0	92.0	10.0	0.0	0.0	38.0	92.0	16.0	12.0
		Overfit	0.0	0.0	0.0	0.0	6.0	0.0	0.0	34.0	6.0	18.0	18.0
	$n = 100$	Underfit	2.0	14.0	2.0	4.0	56.0	88.0	96.0	18.0	12.0	28.0	68.0
		Fit	96.0	84.0	96.0	94.0	26.0	12.0	4.0	36.0	40.0	32.0	18.0
		Overfit	2.0	2.0	2.0	2.0	18.0	0.0	0.0	46.0	48.0	40.0	14.0
	$n = 500$	Underfit	4.0	4.0	4.0	4.0	34.0	56.0	66.0	18.0	6.0	24.0	48.0
		Fit	94.0	94.0	94.0	94.0	32.0	44.0	34.0	32.0	48.0	34.0	36.0
		Overfit	2.0	2.0	2.0	2.0	34.0	0.0	0.0	50.0	46.0	42.0	16.0
$R = 3$ $C = 2$	$n = 50$	Underfit	4.0	6.0	6.0	6.0	100.0	86.0	92.0	64.0	16.0	96.0	92.0
		Fit	96.0	94.0	94.0	94.0	0.0	14.0	8.0	34.0	42.0	4.0	6.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	42.0	0.0	2.0
	$n = 100$	Underfit	2.0	46.0	2.0	2.0	98.0	90.0	92.0	92.0	12.0	98.0	96.0
		Fit	98.0	54.0	98.0	98.0	2.0	10.0	8.0	8.0	28.0	2.0	4.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	60.0	0.0	0.0
	$n = 500$	Underfit	2.0	18.0	2.0	2.0	100.0	90.0	96.0	98.0	18.0	100.0	100.0
		Fit	98.0	82.0	98.0	98.0	0.0	10.0	4.0	2.0	32.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	0.0	0.0
$R = 3$ $C = 3$	$n = 50$	Underfit	74.0	90.0	78.0	86.0	100.0	100.0	100.0	90.0	46.0	100.0	96.0
		Fit	26.0	10.0	22.0	14.0	0.0	0.0	0.0	10.0	32.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.0	0.0	2.0
	$n = 100$	Underfit	62.0	84.0	70.0	72.0	100.0	100.0	100.0	96.0	22.0	100.0	100.0
		Fit	36.0	16.0	30.0	28.0	0.0	0.0	0.0	4.0	48.0	0.0	0.0
		Overfit	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	30.0	0.0	0.0
	$n = 500$	Underfit	6.0	54.0	6.0	6.0	100.0	100.0	100.0	98.0	24.0	100.0	100.0
		Fit	94.0	46.0	94.0	94.0	0.0	0.0	0.0	2.0	44.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.0	0.0	0.0

Notes: All the data is shown in percentage form(%).

Table B.12: Model comparison simulation study results for 11 information criteria. Biclustering ($\mu_k + \phi_k(\alpha_r + \beta_c)$). Scenario 5.

Results			AIC	AIC3	AIC_c	AIC_u	AWE	BIC	CAIC	CLC	NEC	ICL-BIC	$\mathcal{L}$
Overall		Underfit	2.7	13.7	2.8	2.8	62.1	52.2	57.9	40.1	1.5	55.8	63.2
		Fit	92.8	83.3	92.8	92.8	33.4	47.2	41.8	46.6	11.3	36.9	30.9
		Overfit	4.5	3.0	4.3	4.3	4.5	0.7	0.3	13.3	87.2	7.3	5.9
Sample size	$n = 50$	Underfit	4.0	6.0	4.0	4.0	61.3	42.7	52.0	28.0	0.7	51.3	52.0
		Fit	92.7	92.7	92.7	92.7	38.0	57.3	48.0	55.3	12.7	44.7	39.3
		Overfit	3.3	1.3	3.3	3.3	0.7	0.0	0.0	16.7	86.7	4.0	8.7
	$n = 100$	Underfit	2.0	8.7	2.7	2.7	51.3	38.0	43.3	34.0	0.7	46.0	52.7
		Fit	96.0	90.0	96.0	96.0	44.7	62.0	56.7	50.0	4.0	46.7	42.7
		Overfit	2.0	1.3	1.3	1.3	4.0	0.0	0.0	16.0	95.3	7.3	4.7
	$n = 500$	Underfit	1.3	2.7	1.3	1.3	38.0	31.3	41.3	32.7	0.7	33.3	52.7
		Fit	86.0	88.0	86.0	86.0	48.7	66.0	57.3	46.7	1.3	48.7	40.0
		Overfit	12.7	9.3	12.7	12.7	13.3	2.7	1.3	20.7	98.0	18.0	7.3
Number of row clusters	$R = 2$ $C = 2$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	93.3	94.7	93.3	93.3	98.0	98.7	98.7	85.3	0.0	92.0	96.0
		Overfit	6.7	5.3	6.7	6.7	2.0	1.3	1.3	14.7	100.0	8.0	4.0
	$R = 2$ $C = 3$	Underfit	6.0	14.7	6.7	6.7	50.7	72.0	78.7	9.3	0.0	34.0	58.0
		Fit	84.7	80.0	84.7	84.7	33.3	26.7	21.3	53.3	4.0	44.7	25.3
		Overfit	9.3	5.3	8.7	8.7	16.0	1.3	0.0	37.3	96.0	21.3	16.7
	$R = 3$ $C = 2$	Underfit	1.3	2.7	1.3	1.3	100.0	40.0	58.0	85.3	2.0	96.7	99.3
		Fit	96.7	96.0	96.7	96.7	0.0	60.0	42.0	13.3	14.0	3.3	0.7
		Overfit	2.0	1.3	2.0	2.0	0.0	0.0	0.0	1.3	84.0	0.0	0.0
	$R = 3$ $C = 3$	Underfit	3.3	37.3	3.3	3.3	97.7	96.7	95.0	65.7	4.0	92.3	95.3
		Fit	96.7	62.7	96.7	96.7	2.3	3.3	5.0	34.3	27.3	7.7	1.7
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	68.7	0.0	3.0
$R = 2$ $C = 2$	$n = 50$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	92.0	96.0	92.0	92.0	98.0	100.0	100.0	84.0	0.0	94.0	100.0
		Overfit	8.0	4.0	8.0	8.0	2.0	0.0	0.0	16.0	100.0	6.0	0.0
	$n = 100$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	98.0	98.0	98.0	98.0	100.0	100.0	100.0	92.0	0.0	96.0	94.0
		Overfit	2.0	2.0	2.0	2.0	0.0	0.0	0.0	8.0	100.0	4.0	6.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Fit	90.0	90.0	90.0	90.0	96.0	96.0	96.0	80.0	0.0	86.0	94.0
		Overfit	10.0	10.0	10.0	10.0	4.0	4.0	4.0	20.0	100.0	14.0	6.0
$R = 2$ $C = 3$	$n = 50$	Underfit	8.0	12.0	8.0	8.0	84.0	82.0	88.0	14.0	0.0	62.0	56.0
		Fit	92.0	88.0	92.0	92.0	16.0	18.0	12.0	54.0	4.0	32.0	18.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.0	96.0	6.0	26.0
	$n = 100$	Underfit	6.0	24.0	8.0	8.0	54.0	72.0	78.0	14.0	0.0	38.0	60.0
		Fit	90.0	74.0	90.0	90.0	34.0	28.0	22.0	48.0	8.0	44.0	32.0
		Overfit	4.0	2.0	2.0	2.0	12.0	0.0	0.0	38.0	92.0	18.0	8.0
	$n = 500$	Underfit	4.0	8.0	4.0	4.0	14.0	62.0	70.0	0.0	0.0	2.0	58.0
		Fit	72.0	78.0	72.0	72.0	50.0	34.0	30.0	58.0	0.0	58.0	26.0
		Overfit	24.0	14.0	24.0	24.0	36.0	4.0	0.0	42.0	100.0	40.0	16.0
$R = 3$ $C = 2$	$n = 50$	Underfit	4.0	6.0	4.0	4.0	100.0	46.0	68.0	70.0	2.0	92.0	100.0
		Fit	94.0	94.0	94.0	94.0	0.0	54.0	32.0	28.0	34.0	8.0	0.0
		Overfit	2.0	0.0	2.0	2.0	0.0	0.0	0.0	2.0	64.0	0.0	0.0
	$n = 100$	Underfit	0.0	2.0	0.0	0.0	100.0	42.0	52.0	88.0	2.0	100.0	98.0
		Fit	100.0	98.0	100.0	100.0	0.0	58.0	48.0	10.0	4.0	0.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	94.0	0.0	0.0
	$n = 500$	Underfit	0.0	0.0	0.0	0.0	100.0	32.0	54.0	98.0	2.0	98.0	100.0
		Fit	96.0	96.0	96.0	96.0	0.0	68.0	46.0	2.0	4.0	2.0	0.0
		Overfit	4.0	4.0	4.0	4.0	0.0	0.0	0.0	0.0	94.0	0.0	0.0
$R = 3$ $C = 3$	$n = 50$	Underfit	6.0	48.0	6.0	6.0	100.0	100.0	100.0	72.0	4.0	100.0	98.0
		Fit	94.0	52.0	94.0	94.0	0.0	0.0	0.0	28.0	20.0	0.0	0.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	76.0	0.0	2.0
	$n = 100$	Underfit	3.0	36.0	3.0	3.0	98.0	97.0	94.0	67.0	4.0	91.0	94.0
		Fit	97.0	64.0	97.0	97.0	2.0	3.0	6.0	33.0	27.0	9.0	3.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	69.0	0.0	3.0
	$n = 500$	Underfit	1.0	28.0	1.0	1.0	95.0	93.0	91.0	58.0	4.0	86.0	94.0
		Fit	99.0	72.0	99.0	99.0	5.0	7.0	9.0	42.0	35.0	14.0	2.0
		Overfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	61.0	0.0	4.0

Notes: All the data is shown in percentage form(%).

Table B.13: Full city-knowledge questionnaire. Informants rated the following 20 questions for either Irvine, New York, or Miami.

Questions
Q1. Rate the amount of rain experienced during the fall.
Q2. Rate the amount of snow experienced during the winter.
Q3. Rate the level of humidity in the summer.
Q4. Rate the general wind factor during the fall.
Q5. Rate how cold it is during the winter.
Q6. Rate how hot it is during the summer.
Q7. Rate the range of temperatures experienced across the year.
Q8. Rate the amount of people that use public transportation as the primary mode of transport.
Q9. Rate the amount of crime that occurs.
Q10. Rate the amount of ethnic/racial diversity.
Q11. Rate how liberally minded the general population is.
Q12. Rate how much "nightlife" the city has.
Q13. Rate the population density of the city.
Q14. Rate how close the ocean is.
Q15. Rate how modernized the city is.
Q16. Rate the air quality (smog level) of the city.
Q17. Rate the cleanliness of the city.
Q18. Rate how well-known the city is compared to other cities in the state.
Q19. Rate the cost of living in the city.
Q20. Rate the amount of homeless people living in the city.

Table B.14: Full city-knowledge questionnaire responses (Anders and Batchelder, 2013). 83 respondents (rows) answering 20 knowledge questions (Q1-Q20) pertaining to particular city on a 7-point scale categories.

ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	True City
1	2	1	3	2	2	3	1	3	1	5	3	3	3	6	6	4	6	4	5	1	Irvine
2	5	7	4	5	7	5	4	5	5	4	4	5	6	7	5	5	5	7	4	5	New York
3	4	5	7	3	7	7	6	7	6	6	5	7	7	3	6	7	1	7	7	5	New York
4	5	1	6	5	3	6	2	2	3	4	5	6	5	7	6	4	5	7	6	4	Miami
5	4	2	1	7	7	3	6	5	5	7	6	6	7	5	6	3	3	7	6	5	New York
6	2	1	4	5	5	5	5	2	2	3	4	3	4	1	5	3	6	5	6	1	Irvine
7	4	1	2	5	6	6	4	2	1	4	4	2	5	6	7	6	7	4	6	2	Irvine
8	2	1	6	4	3	7	5	6	5	5	4	6	6	7	6	6	3	6	5	5	Miami
9	6	3	7	5	3	7	5	4	5	7	6	7	6	6	7	6	5	7	6	7	Miami
10	2	1	3	2	3	6	4	2	1	4	4	3	4	6	7	3	6	6	7	2	Irvine
11	2	1	4	2	4	4	4	2	1	3	6	2	4	7	7	5	7	2	7	1	Irvine
12	2	1	7	3	4	7	4	5	6	6	5	7	5	7	6	4	3	5	6	5	Miami
13	2	1	5	3	3	6	4	2	1	3	4	1	2	6	5	3	6	2	6	1	Irvine
14	4	1	7	3	3	5	2	3	6	7	4	6	5	6	6	2	4	7	5	5	Miami
15	2	1	2	4	2	6	4	2	1	6	4	2	4	6	5	6	6	3	4	1	Irvine
16	3	1	2	2	3	5	2	1	1	2	2	2	3	6	5	3	6	4	6	1	Irvine
17	3	2	2	5	6	6	5	3	2	6	5	2	3	6	6	4	5	5	5	2	Miami
18	6	6	6	4	5	5	4	7	5	6	6	7	6	6	7	3	4	7	7	5	New York
19	2	1	4	5	6	7	4	6	1	5	6	1	5	5	6	2	7	4	6	1	Irvine
20	6	1	7	4	2	7	4	1	5	5	4	7	7	7	6	7	1	7	5	6	Miami
21	5	6	7	5	7	6	7	7	7	6	7	7	7	6	7	7	3	7	7	6	New York
22	6	7	6	6	7	4	6	7	6	6	6	6	5	5	6	3	4	5	7	5	New York
23	4	1	3	4	6	5	5	7	1	5	5	3	5	5	6	1	7	5	7	1	Irvine
24	5	1	7	4	2	7	5	6	6	6	7	7	6	7	5	4	5	7	4	5	Miami
25	4	4	5	3	3	7	5	5	4	6	4	7	6	7	6	4	5	7	3	5	Miami
26	4	4	5	4	6	5	6	7	6	6	6	7	7	6	6	7	3	7	7	7	New York
27	2	1	5	5	6	4	5	2	2	6	3	2	2	6	4	6	6	4	6	2	Irvine
28	1	1	6	6	2	4	3	1	1	5	1	2	3	7	2	6	6	3	5	2	Irvine
29	3	1	3	5	5	5	3	4	1	3	4	4	4	6	5	2	6	5	5	1	Irvine
30	3	1	1	4	4	6	5	6	2	3	4	1	3	6	7	7	6	3	6	1	Irvine
31	6	1	7	5	5	7	5	4	6	4	4	7	6	7	5	2	1	6	5	6	Miami
32	3	7	2	3	6	3	4	7	5	5	7	7	6	6	7	1	2	7	7	7	New York
33	3	2	6	4	5	6	4	5	2	4	4	5	5	6	4	3	3	4	4	4	Miami
34	5	1	3	6	1	6	4	3	2	6	7	7	7	7	7	4	6	7	4	6	Miami
35	5	1	7	4	3	7	5	3	4	5	4	6	5	6	5	5	4	6	6	4	Miami
36	5	4	3	5	6	5	4	7	4	5	5	6	6	4	6	6	3	7	6	5	New York
37	4	1	3	5	5	5	4	3	2	5	3	3	4	4	5	6	6	4	6	1	Irvine
38	2	2	6	4	4	7	4	3	3	4	5	6	3	7	5	6	5	6	5	4	Miami
39	4	6	4	3	6	5	7	7	5	7	6	7	7	7	7	2	4	7	6	6	New York
40	7	1	7	5	1	7	5	2	5	3	4	7	7	7	7	7	2	7	7	6	Miami
41	6	5	7	7	7	7	5	5	5	5	5	7	7	7	7	1	4	5	7	6	Miami
42	3	4	5	6	3	7	3	1	5	6	6	7	4	7	5	4	5	3	4	6	Miami
43	5	4	4	4	6	5	5	7	5	6	5	5	7	6	6	4	4	7	5	5	New York
44	5	1	6	4	5	6	3	2	1	1	1	1	1	7	4	6	7	3	7	1	Irvine
45	3	1	5	3	5	6	6	7	1	4	1	3	2	3	5	5	6	7	6	2	Irvine
46	5	6	6	5	6	5	6	7	5	5	5	7	6	6	7	3	2	6	6	5	New York
47	3	1	5	4	6	7	3	3	6	6	2	7	6	5	5	2	3	7	6	4	Miami
48	6	2	7	4	6	7	6	5	6	5	4	7	6	7	7	3	2	7	6	6	Miami
49	2	3	6	5	4	7	5	4	3	5	4	6	5	7	6	3	4	6	6	2	Irvine
50	2	1	6	4	3	6	4	4	3	7	3	7	6	7	6	4	4	7	6	3	Miami
51	4	5	3	5	6	6	4	5	6	4	6	7	6	7	6	1	3	7	5	6	Miami
52	3	1	3	5	4	5	5	1	1	2	3	1	2	6	5	6	7	4	6	1	Miami
53	2	1	6	3	2	6	4	5	5	4	5	6	5	7	5	4	3	6	5	5	Miami
54	6	6	2	5	7	3	3	7	4	4	6	7	7	6	7	4	4	7	6	4	New York
55	6	2	6	4	4	5	4	3	4	6	6	5	5	6	5	5	5	6	5	4	Miami
56	5	5	5	3	6	6	5	6	5	6	5	5	6	6	6	6	4	7	6	5	New York
57	6	6	5	7	7	6	6	7	5	5	4	2	7	7	7	7	3	7	7	6	New York
58	3	1	6	4	4	6	4	2	2	1	1	1	3	3	2	3	5	3	7	1	Irvine
59	4	2	5	4	3	4	4	4	3	4	4	5	4	6	6	5	5	7	5	4	Miami
60	4	1	5	3	4	6	5	6	2	7	7	2	5	2	6	5	7	4	5	5	Irvine
61	2	1	3	5	6	6	5	2	1	3	4	2	5	6	6	5	6	3	6	2	Irvine
62	6	7	4	5	7	4	4	7	5	6	6	7	5	4	6	1	1	7	6	6	New York
63	6	1	6	4	3	6	5	5	5	4	4	6	6	7	6	5	5	6	6	5	Miami
64	1	1	1	4	6	5	6	2	1	3	4	2	5	4	6	1	7	6	6	1	Irvine
65	6	7	5	5	7	3	5	7	5	7	7	7	7	3	7	7	4	7	7	7	New York
66	4	7	5	3	6	6	7	7	6	7	4	7	5	6	6	2	3	6	3	4	New York
67	1	2	1	4	6	3	6	3	5	6	5	4	5	7	7	5	5	5	6	4	Miami
68	1	1	1	3	3	7	5	2	1	1	2	2	3	6	6	7	7	5	7	2	Irvine
69	2	1	3	3	3	5	5	2	1	2	1	1	2	6	7	7	7	2	6	1	Irvine
70	7	7	7	5	7	7	7	7	7	5	5	7	7	3	7	7	1	7	7	7	New York
71	7	1	7	7	5	7	5	7	7	7	4	7	7	7	7	7	3	7	7	6	Miami
72	4	3	4	4	6	3	4	6	6	3	4	6	6	6	6	6	4	7	5	6	New York
73	3	1	6	5	3	6	4	1	1	3	4	1	5	4	5	4	6	7	7	2	Irvine
74	3	1	4	3	5	3	5	2	1	2	3	2	2	6	7	7	7	4	7	2	Irvine
75	3	1	4	5	5	6	5	2	1	4	5	2	4	6	6	7	7	5	6	2	Irvine
76	3	1	7	4	3	7	5	1	5	5	5	6	6	7	7	6	3	7	7	5	Miami
77	4	4	6	5	7	3	5	7	5	7	5	7	7	5	5	7	5	7	7	6	New York
78	6	7	5	5	7	4	7	7	6	5	6	7	7	3	6	2	2	7	7	7	New York
79	5	7	2	6	7	4	6	7	5	6	7	7	5	2	6	3	2	6	6	5	New York
80	1	3	6	5	6	3	4	3	2	3	4	2	4	5	6	1	7	6	6	2	Irvine
81	1	1	6	6	4	6	6	2	2	3	4	3	5	7	5	4	6	5	6	1	Irvine
82	2	1	3	5	4	6	4	3	1	4	1	1	3	6	7	6	7	4	6	1	Irvine
83	5	6	7	4	7	7	5	7	6	7	7	7	7	5	6	7	2	7	7	7	New York

Appendix C

Data Applications. Results EM Algorithm

C.1 Simulation Study. Other Scenarios

Tables C.1 and C.2 summarise the results for the simulation scenarios (Section 4.1) including the interaction factors for row clustering and biclustering version respectively. Figures C.1 and C.2 show how the precision of the estimates of the score parameters $\widehat{\phi}_2$ and $\widehat{\phi}_3$ depends on sample size n in the case with $R = 2$ row groups with $C = 2$ column groups and biclustering with $R = 2$ and $C = 2$ row and column groups, respectively. Tables C.3, C.4 and C.5 show two particular scenarios described in Section 4.1 for the row clustering, column clustering and biclustering respectively.

C.2 Applied Statistics Course Feedback Forms

The list of questions are shown in Table C.6 and the data set with the responses of 70 students over 10 questions giving feedback about a second year Applied Statistics course is given in Table C.7.

C.3 Tree Presences in Great Smoky Mountains

The data set of R. H. Whittaker's study of vegetation of the Great Smoky Mountains (Whittaker, 1956, Table 3) is shown in Table C.8. The data set consists of the

Table C.1: Simulation study (Section 4.1). Estimated score parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j} + \boldsymbol{\gamma_{rj}})$. MLEs and their standard errors from the score and row membership parameters ($\{\phi_k\}$,$\{\pi_r\}$) for different number of row clusters R and sample sizes n are shown.

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

distribution of 41 different tree species along 12 different site stations located at altitudes between 3500 and 4500 ft and sorted by moisture level (wetter to drier). Table C.9 summarises the suite of fitted models for this data set. For each model, the information criteria AIC, AIC_c, BIC and ICL-BIC were computed.

C.4 Spider Data

The spider abundance data set (Van der Aart and Smeenk-Enserink, 1974) shows the distribution of 12 different spider species across 28 different sites. The original count data is shown in Table C.10 and the categorised ordinal data set following eq. (4.5) is shown in Table C.11. Table C.12 summarises the suite of fitted models

Table C.2: Simulation study (Section 4.1). Estimated score parameters for stereotype model including biclustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_c} + \boldsymbol{\gamma_{rc}})$. MLEs and their standard errors from the score, row and column membership parameters ($\{\phi_k\}$,$\{\pi_r\}$,$\{\kappa_c\}$) for different number of row and column clusters R and C and sample sizes n are shown.

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

for this data set. For each model, the information criteria AIC, AIC_c, BIC and ICL-BIC were computed.

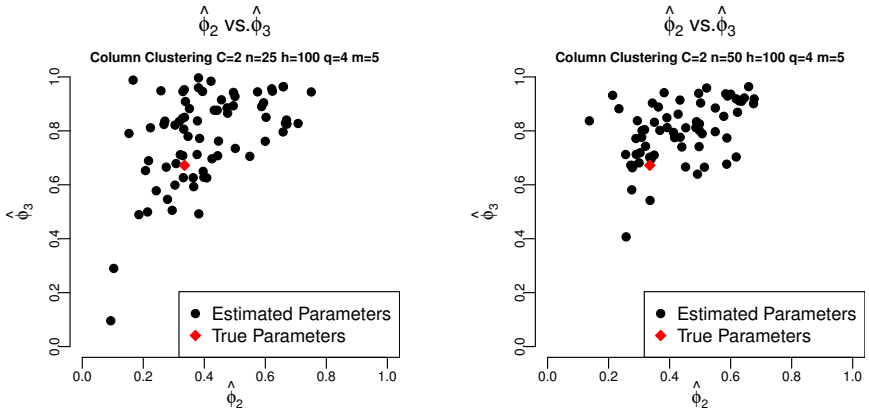

Figure C.1: ***Simulation study (Section 4.1)***: *Convergence of $\widehat{\phi}_2$ and $\widehat{\phi}_3$ for the stereotype model including column clustering $(\alpha_i + \beta_c)$ with $C = 2$ column clusters. n, h, q, m describe the sample size, the number of replicates, the number of categories and the number of covariates respectively. The diamond point represents the true value of the parameter.*

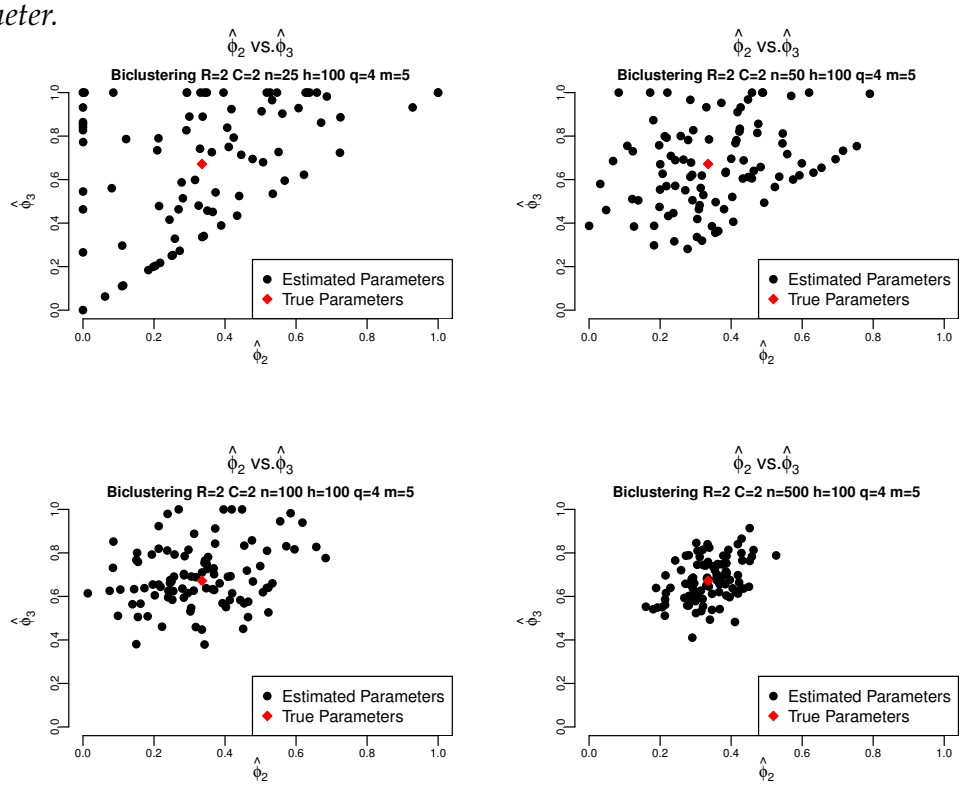

Figure C.2: ***Simulation study (Section 4.1)***: *Convergence of $\widehat{\phi}_2$ and $\widehat{\phi}_3$ for the stereotype model including biclustering $(\alpha_r + \beta_c)$ with $R = 2$ row and $C = 2$ column clusters. n, h, q, m describe the sample size, the number of replicates, the number of categories and the number of covariates respectively. The diamond point represents the true value of the parameter.*

Table C.3: Simulation study (Section 4.1). Estimated score parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j} + \boldsymbol{\gamma_{rj}})$ when $\phi_2 = \phi_3$ or π_2 is small. MLEs and their standard errors from the score and row membership parameters ($\{\phi_k\}$,$\{\pi_r\}$) for different number of row clusters R and sample sizes n are shown.

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

Table C.4: Simulation study (Section 4.1). Estimated score parameters for stereotype model including column clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_i} + \boldsymbol{\beta_c} + \boldsymbol{\gamma_{ic}})$ when $\phi_2 = \phi_3$ or κ_2 is small and $m = 15$. MLEs and their standard errors from the score and column membership parameters ($\{\phi_k\}$,$\{\kappa_c\}$) for different number of column clusters C, number of columns m and sample sizes n are shown.

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

Table C.5: Simulation study (Section 4.1). Estimated score parameters for stereotype model including biclustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_c} + \boldsymbol{\gamma_{rc}})$ when $\phi_2 = \phi_3$ or π_2 and κ_2 are small. MLEs and their standard errors from the score, row and column membership parameters ($\{\phi_k\}$,$\{\pi_r\}$,$\{\kappa_c\}$) for different number of row and column clusters R and C and sample sizes n are shown.

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

Table C.6: List of 10 questions of Applied Statistics course feedback forms data set. Each question was written so that "agree" indicates a positive view of the course.

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

Table C.7: Applied Statistics feedback forms data set. 70 students (rows), 10 questions (Q1-Q10) and 3 categories for each question: "disagree" (coded as 1), "neither agree or disagree" (coded as 2) and "agree" (coded as 3).

Students ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	1	1	1	1	2	2	2	2	1	1
2	1	1	1	1	3	3	3	3	1	1
3	1	1	1	2	1	2	3	2	2	2
4	2	2	1	1	2	2	3	3	2	2
5	1	1	1	1	2	3	3	3	1	3
6	3	2	1	2	1	3	3	3	2	3
7	1	1	1	3	2	2	3	1	1	1
8	1	1	1	1	1	2	2	1	2	1
9	2	1	1	1	2	3	2	3	1	2
10	1	1	1	1	2	2	2	2	2	2
11	1	1	1	1	1	1	1	1	1	1
12	2	1	1	3	3	3	3	2	1	3
13	2	1	1	2	3	3	3	1	1	2
14	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	2	1	1
16	1	1	1	2	2	1	3	2	1	2
17	1	1	1	2	2	2	2	2	1	2
18	1	1	1	1	1	2	2	1	1	2
19	1	1	1	1	2	2	1	1	1	1
20	2	2	1	2	1	2	2	2	1	2
21	1	1	1	1	1	2	2	1	1	1
22	3	3	1	1	1	1	2	1	1	1
23	1	1	1	1	1	2	2	1	1	1
24	1	1	1	1	1	2	2	2	1	1
25	1	1	1	1	1	2	2	2	1	1
26	1	1	2	2	1	1	2	3	1	2
27	1	1	1	1	1	1	2	1	1	1
28	1	1	1	1	1	1	1	1	1	1
29	1	1	1	2	2	2	2	2	2	2
30	1	1	1	2	2	3	3	2	1	2
31	1	1	1	1	1	1	1	2	1	1
32	1	1	1	1	1	2	1	1	1	1
33	2	1	1	2	1	2	2	2	1	2
34	1	1	1	2	2	2	2	2	2	2
35	1	2	1	1	2	2	2	3	1	2
36	1	1	1	2	1	2	1	2	2	2
37	1	1	1	2	2	2	2	2	1	2
38	1	1	1	2	2	2	2	1	1	1
39	1	1	1	1	1	1	1	2	1	2
40	1	1	2	2	1	3	3	1	2	2
41	1	1	1	1	1	3	1	3	1	1
42	1	1	1	1	2	2	2	1	1	1
43	1	1	1	2	2	3	3	3	2	2
44	2	1	1	2	2	3	2	3	1	1
45	3	1	1	2	3	3	3	1	1	3
46	1	1	1	1	2	2	2	2	1	2
47	1	1	1	1	1	2	2	3	1	1
48	1	1	1	2	1	1	2	1	1	2
49	1	1	1	2	1	2	1	2	1	1
50	1	1	1	1	2	2	2	3	2	2
51	1	1	1	1	2	3	3	2	2	2
52	1	1	1	1	1	2	2	3	2	1
53	1	1	1	1	1	1	1	1	1	1
54	2	3	1	3	3	3	3	3	3	2
55	1	1	1	2	1	1	1	2	1	1
56	1	1	1	1	1	1	1	1	1	1
57	1	1	1	2	2	3	2	2	1	1
58	1	1	1	1	1	2	2	2	1	1
59	1	1	1	1	1	1	2	1	2	1
60	2	1	1	2	3	3	3	3	2	2
61	1	1	1	1	1	1	2	1	1	2
62	1	1	1	1	2	2	2	1	1	1
63	1	1	1	2	1	2	2	3	1	2
64	1	1	1	1	3	3	3	3	1	1
65	1	1	1	1	1	1	3	1	1	1
66	1	1	1	2	1	2	2	2	1	2
67	3	1	1	1	1	1	1	3	1	3
68	1	1	1	1	2	2	2	3	1	1
69	1	3	1	1	2	2	2	2	1	1
70	3	1	1	3	3	3	1	3	1	2

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

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

Table C.9: Suite of models fitted for R.H. Whittaker's study of vegetation. For each information criterion, the best model in each group (no clustering, row clustering, column clustering and biclustering) is shown in boldface.

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

Table C.10: Spider data set (Van der Aart and Smeenk-Enserink, 1974). Abundances of $m = 12$ species of spiders recorded at $n = 28$ sites.

Sites	Species of spiders											
	Alopacce	*Alopcune*	*Alopfabr*	*Arctlute*	*Arctperi*	*Auloalbi*	*Pardlugu*	*Pardmont*	*Pardnigr*	*Pardpull*	*Trocterr*	*Zoraspin*
1	25	10	0	0	0	4	0	60	12	45	57	4
2	0	2	0	0	0	30	1	1	15	37	65	9
3	15	20	2	2	0	9	1	29	18	45	66	1
4	2	6	0	1	0	24	1	7	29	94	86	25
5	1	20	0	2	0	9	1	2	135	76	91	17
6	0	6	0	6	0	6	0	11	27	24	63	34
7	2	7	0	12	0	16	1	30	89	105	118	16
8	0	11	0	0	0	7	55	2	2	1	30	3
9	1	1	0	0	0	0	0	26	1	1	2	0
10	3	0	1	0	0	0	0	22	0	0	1	0
11	15	1	2	0	0	1	0	95	0	1	4	0
12	16	13	0	0	0	0	0	96	1	8	13	0
13	3	43	1	2	0	18	1	24	53	72	97	22
14	0	2	0	1	0	4	3	14	15	72	94	32
15	0	0	0	0	0	0	6	0	0	0	25	3
16	0	3	0	0	0	0	6	0	2	0	28	4
17	0	0	0	0	0	0	2	0	0	0	23	2
18	0	1	0	0	0	0	5	0	0	0	25	0
19	0	1	0	0	0	0	12	0	1	0	22	3
20	0	2	0	0	0	0	13	0	0	0	22	2
21	0	1	0	0	0	0	16	1	0	1	18	2
22	7	0	16	0	4	0	0	2	0	0	1	0
23	17	0	15	0	7	0	2	6	0	0	1	0
24	11	0	20	0	5	0	0	3	0	0	0	0
25	9	1	9	0	0	2	1	11	6	0	16	6
26	3	0	6	0	18	0	0	0	0	0	1	0
27	29	0	11	0	4	0	0	1	0	0	0	0
28	15	0	14	0	1	0	0	6	0	0	2	0

Table C.11: Spider data set (Van der Aart and Smeenk-Enserink, 1974). Abundances of $n = 12$ species of spiders recorded at $m = 28$ sites grouped in ordinal categories (as described in eq. (4.5)).

Sites	Species of spiders											
	Alopacce	*Alopcune*	*Alopfabr*	*Arctlute*	*Arctperi*	*Auloalbi*	*Pardlugu*	*Pardmont*	*Pardnigr*	*Pardpull*	*Trocterr*	*Zoraspin*
1	2	2	0	0	0	1	0	3	2	3	3	1
2	0	2	0	0	0	3	1	1	2	3	3	2
3	2	3	1	1	0	2	1	3	2	3	3	1
4	1	2	0	1	0	2	1	2	3	3	3	2
5	1	2	0	1	0	2	1	1	3	3	3	2
6	0	1	0	1	0	1	0	2	3	2	3	3
7	1	1	0	2	0	2	1	2	3	3	3	2
8	0	3	0	0	0	2	3	1	1	1	3	2
9	1	1	0	0	0	0	0	3	1	1	3	0
10	2	0	1	0	0	0	0	3	0	0	1	0
11	3	1	2	0	0	1	0	3	0	1	2	0
12	3	2	0	0	0	0	0	3	1	1	2	0
13	2	3	1	1	0	2	1	2	3	3	3	2
14	0	1	0	1	0	2	1	2	2	3	3	3
15	0	0	0	0	0	0	2	0	0	0	3	1
16	0	1	0	0	0	0	3	0	1	0	3	2
17	0	0	0	0	0	0	1	0	0	0	3	1
18	0	1	0	0	0	0	2	0	0	0	3	0
19	0	1	0	0	0	0	3	0	1	0	3	2
20	0	1	0	0	0	0	2	0	0	0	3	1
21	0	1	0	0	0	0	3	1	0	1	3	2
22	3	0	3	0	2	0	0	1	0	0	1	0
23	3	0	3	0	2	0	1	2	0	0	1	0
24	2	0	3	0	2	0	0	1	0	0	0	0
25	2	1	2	0	0	1	1	3	2	0	3	2
26	2	0	2	0	3	0	0	0	0	0	1	0
27	3	0	2	0	2	0	0	1	0	0	0	0
28	3	0	3	0	1	0	0	2	0	0	1	0

Table C.12: Suite of models fitted for spider data set (Van der Aart and Smeenk-Enserink, 1974). For each information criterion, the best model in each group (no clustering, row clustering, column clustering and biclustering) is shown in boldface.

Model		R	C	npar	AIC	AIC_c	BIC	ICL-BIC
Null effects	$\mu_k + \phi_k$	1	1	5	441.63	441.81	**460.71**	**460.71**
Row effects	$\mu_k + \phi_k\alpha_i$	n	1	16	428.81	430.52	489.89	489.89
Column effects	$\mu_k + \phi_k\beta_j$	1	m	32	463.85	470.82	586.00	586.00
Main effects	$\mu_k + \phi_k(\alpha_i + \beta_j)$	n	m	43	**422.54**	**421.50**	547.67	547.67
Row Clustering	$\mu_k + \phi_k\alpha_r$	2	1	7	415.70	**416.04**	**442.42**	**442.49**
		3	1	9	419.42	419.97	453.77	470.37
		4	1	11	423.36	424.17	465.35	481.86
		5	1	13	427.40	428.53	477.02	496.25
		6	1	15	430.96	432.46	488.22	488.24
	$\mu_k + \phi_k(\alpha_r + \beta_j)$	2	m	34	431.02	438.92	560.80	572.87
		3	n	20	435.91	444.82	573.33	594.32
		4	n	22	439.57	449.55	584.62	593.90
		5	n	24	443.91	455.03	596.60	599.43
		6	n	26	447.69	460.02	608.01	618.21
	$\mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj})$	2	m	61	**406.22**	423.83	629.06	639.08
		3	n	42	424.71	491.57	668.25	776.26
		4	n	55	426.25	558.47	680.49	681.49
		5	n	68	549.95	585.80	681.88	684.89
		6	n	81	531.77	630.58	707.40	717.40
Column Clustering	$\mu_k + \phi_k\beta_c$	1	2	7	412.46	412.81	**439.18**	**463.05**
		1	3	9	418.12	418.67	452.47	482.00
		1	4	11	421.90	422.71	463.89	515.37
		1	5	13	426.43	427.56	476.06	507.19
		1	6	15	429.96	431.46	487.22	547.28
	$\mu_k + \phi_k(\alpha_i + \beta_c)$	n	2	18	410.13	415.81	520.82	526.18
		$\boldsymbol{n}$	**3**	**20**	**397.28**	**409.28**	561.54	565.73
		n	4	22	401.23	413.55	607.22	609.89
		n	5	24	412.15	447.29	671.71	675.77
		n	6	26	460.91	513.21	770.10	772.98
	$\mu_k + \phi_k(\alpha_i + \beta_c + \gamma_{ic})$	n	2	29	534.06	538.66	664.21	669.38
		n	3	42	436.57	439.24	512.92	542.04
		n	4	55	440.43	443.66	524.41	549.82
		n	5	68	444.03	447.89	535.64	554.73
		n	6	81	450.14	454.68	549.38	595.48
Biclustering	$\mu_k + \phi_k(\alpha_r + \beta_c)$	2	2	9	421.76	422.31	456.11	498.31
		2	3	11	**419.64**	**420.20**	**454.00**	490.75
		2	4	13	425.74	426.88	475.37	549.88
		2	5	15	431.31	432.81	488.56	572.19
		3	2	11	423.22	424.03	465.20	517.86
		3	3	13	476.66	477.79	501.77	526.29
		3	4	15	439.87	441.37	497.13	522.80
		3	5	17	435.21	437.13	500.10	567.88
		4	2	13	482.98	484.11	492.13	532.60
		4	3	15	433.70	435.20	490.96	550.30
		4	4	17	435.22	437.14	500.11	571.15
		4	5	19	464.04	466.44	536.56	568.45
	$\mu_k + \phi_k(\alpha_r + \beta_c + \gamma_{rc})$	2	2	10	427.97	429.10	477.59	527.43
		2	3	13	422.00	422.68	460.17	**486.88**
		2	4	16	434.39	436.09	495.46	520.85
		2	5	19	438.61	441.01	511.13	538.56
		3	2	13	497.76	498.89	505.27	547.38
		3	3	17	433.91	435.84	498.80	540.76
		3	4	21	441.89	444.83	522.05	559.23
		3	5	25	453.08	457.27	548.50	615.81
		4	2	16	445.85	447.55	506.92	528.75
		4	3	21	448.82	451.76	528.98	538.18
		4	4	26	468.71	473.25	567.95	622.25
		4	5	31	530.60	537.12	619.79	648.93

Appendix D

Spaced Mosaic Plots. R function

In this section we describe the **R** function to fit a spaced mosaic plot. The purpose of showing the function here is to use it independently of any package. This function will be included within an **R** package and its structure might change (e.g. defining a **R** class object related to the mosaic). In the meantime, you can e-mail the corresponding author (D. Fernández - `daniel.fernandez@msor.vuw.ac.nz`) to obtain this function.

The description of the **R** function we have developed is:

`spaced.mosaic.plot` *Draw spaced mosaic plots for clustering ordinal data*

Description

The function `spaced.mosaic.plot` computes a spaced mosaic plot of a given ordinal data, clustering structure and fitted score parameters.

Usage

```
spaced.mosaic.plot(mdata, phi, R, ClusterRow, labels)
```

Arguments

`mdata`	numeric matrix containing the ordinal data set.
`phi`	numeric vector representing the fitted score parameters ($\{\phi_k\}$)

from the ordinal stereotype model.

`R` integer value specifying the number of row clusters.

`ClusterRow` an integer vector with the allocated cluster to each row.

`labels` (optional) a list comprising 4 items:

- `categ` contains the labels for the ordinal categories.
- `cluster` contains the labels for the clusters.
- `row` contains the labels for the data rows.
- `col` contains the labels for the data columns.

Value(s)

The function returns a data frequency table with `R` rows and one column for each category. In addition, three `pdf` files are generated in the working directory with the overall distribution (MosaicPlot_withoutRowCluster.pdf), the row clustering structure (MosaicPlot_R=*R*.pdf) and the inclusion of the space between adjacent ordinal categories (MosaicPlot_SPACING_R=*R*.pdf).

Author(s)

Daniel Fernández

References

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

See Also

mosaicplot

Example

```
library(grid)
library(vcd)
#Score parameters
phi <- c(0,0.5,0.7,1)

#Generation of simulated data
q <- length(phi)
n <- 28
m <- 12
R <- 3
labels <- list(categ=c("Disagree", "No Opinion", "Agree",
               "Strongly Agree"), cluster=paste("R",1:R,sep=""),
               row=paste("r",1:n,sep=""),
               col=paste("c",1:m,sep=""))
y.mat <- matrix(NA,n,m)
for(i in 1:n) for (j in 1:m) y.mat[i,j] <- sample(1:q,1,prob=c())

ClusterRowY <- array(NA,n)
for (i in 1:n) ClusterRowY[i] <- sample(1:R,1)
rownames(ClusterRowY) <- labels$row

#Generate spaced mosaic plot
spaced.mosaic.plot(y.mat, phi, R, ClusterRowY, labels)
```

```
  Col
Row  Disagree No Opinion Agree Strongly Agree
  R1       31         41    29            31
  R2       34         32    27            27
  R3       21         21    21            21
```

Appendix E

Metropolis-Hastings. Definitions

E.1 Degenerate Normal Distribution

When we define a n-multivariate normal distribution, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, we assume that the covariate matrix $\boldsymbol{\Sigma}$ is a strictly positive definite and symmetric matrix and the support of the distribution is the full space $\mathbb{R}^n$. We may however be interested in a distribution confined to a hyperplane of lower dimension contained in $\mathbb{R}^n$. We refer to this type of distribution as a degenerate normal distribution with mean and covariance matrix $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively (DegenNormal $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). However, $\boldsymbol{\Sigma}$ is no longer positive definite. In the following two sections we formally define this distribution for the one and two-dimensional cases in the particular case when we have equal mean $\boldsymbol{\mu}$ and impose the constraint $\sum_{i=1}^{n} x_i = n\mu$ on the random variable $\boldsymbol{x}$.

E.1.1 One-Dimensional

If a $n \times 1$ vector $\boldsymbol{x}$ follows an one-dimensional degenerate normal with mean $\mu \underline{\mathbf{1}}_n$ and variance-covariance matrix $\boldsymbol{\Sigma}$ then we write

$$\boldsymbol{x} \sim \text{DegenNormal}(n; \mu \underline{\mathbf{1}}_n, \boldsymbol{\Sigma}),$$

where

$$\Sigma_{ij} = \begin{cases} -\frac{\sigma^2}{n} & \text{if } i \neq j \\ \\ \sigma^2 \left(1 - \frac{1}{n}\right) & \text{if } i = j \end{cases},$$

which is confined to the hyperplane $\sum_{i=1}^{n} x_i = n\mu$. We use a delta function to apply this constraint and then the formulation of the density of this distribution is given by

$$f_{\mathrm{Deg}\mathcal{N}}(\boldsymbol{x}|n,\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}(n-1)} n^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \times \delta\left(\sum_{i=1}^{n} x_i - n\mu\right). \tag{E.1}$$

Any individual component of $\boldsymbol{x}$ has a marginal normal distribution:

$$x_i \sim \mathcal{N}\left(\mu, \frac{n-1}{n}\sigma^2\right) \qquad i = 1,\ldots,n. \tag{E.2}$$

Any two different components x_i and x_j ($i \neq j$) are negatively correlated with

$$\mathrm{Cov}(x_i, x_j) = -\sigma^2/n. \tag{E.3}$$

Proof eq. (E.2). We define n independent and identically distributed random standard normal variables: $z_i \sim \mathcal{N}(0,1)$ $(i = 1, \ldots, n)$.
Set $x_i = \sigma(z_i - \overline{z}) + \mu$ as a linear combination of z_i $(i = 1, \ldots, n)$,
where $\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ and the hyperplane $\sum_{i=1}^{n} x_i$ is

$$\sum_{i=1}^{n} x_i = \sigma(n\overline{z} - n\overline{z}) + n\mu = n\mu.$$

In addition,

$$x_i = \sigma(z_i - \overline{z}) + \mu = \sigma\left(z_i - \frac{1}{n}\sum_{i=1}^{n} z_i\right) + \mu = \sigma\left(z_i - \frac{1}{n}z_i - \frac{1}{n}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_\ell\right) + \mu$$
$$= \sigma\left(1 - \frac{1}{n}\right) z_i - \sigma\frac{1}{n}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_\ell + \mu.$$

Therefore, x_i is normally distributed with

$$E[x_i] = E\left[\sigma\left(1 - \frac{1}{n}\right) z_i\right] - E\left[\sigma\frac{1}{n}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_\ell\right] + E\left[\mu\right]$$
$$= \sigma\left(1 - \frac{1}{n}\right) E[z_i] - \frac{\sigma(n-1)}{n}E[z_i] + \mu = \mu \qquad \text{and}$$

$$V[x_i] = V\left[\sigma\left(1 - \frac{1}{n}\right) z_i\right] + V\left[\sigma\frac{1}{n}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_\ell\right] + V\left[\mu\right]$$
$$= \sigma^2\left(1 - \frac{1}{n}\right)^2 V[z_i] + \frac{\sigma^2(n-1)}{n^2}V[z_i] = \left(\frac{n-1}{n}\right)^2 \sigma^2 + \frac{n-1}{n^2}\sigma^2 =$$
$$= \frac{n-1}{n^2}\left(\sigma^2(n-1) + \sigma^2\right) = \frac{n-1}{n^2}\sigma^2 n = \frac{n-1}{n}\sigma^2.$$

■

Proof eq. (E.3). We define n independent and identically distributed random standard normal variables: $z_i \sim \mathcal{N}(0,1)$ $(i = 1, \ldots, n)$.
Set $x_i = \sigma(z_i - \overline{z}) + \mu$ as a linear combination of z_i, where $\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ and $E[z_i^2] = V[z_i] + (E[z_i])^2 = 1$ $(i = 1, \ldots, n)$.
Thus, the covariance between any two different individuals components is formulated as:

$$
\begin{aligned}
\text{Cov}(x_i, x_j) &= E[(x_i - \overline{x})(x_j - \overline{x})] = E\left[(\sigma(z_i - \overline{z}) + \mu - \mu)\left(\sigma(z_i - \overline{z}) + \mu - \mu\right)\right] \\
&= \sigma^2 E\left[\left(z_i - \frac{1}{n}z_i - \frac{1}{n}z_j - \frac{1}{n}\sum_{\substack{\ell=1 \\ \ell \notin \{i,j\}}}^{n} z_\ell\right)\left(z_j - \frac{1}{n}z_i - \frac{1}{n}z_j - \frac{1}{n}\sum_{\substack{\ell'=1 \\ \ell' \notin \{i,j\}}}^{n} z_{\ell'}\right)\right] \\
&= \sigma^2 E\left[\left(\left(1 - \frac{1}{n}\right)z_i - \frac{1}{n}z_j - \frac{1}{n}\sum_{\substack{\ell=1 \\ \ell \notin \{i,j\}}}^{n} z_\ell\right)\left(\left(1 - \frac{1}{n}\right)z_j - \frac{1}{n}z_i - \frac{1}{n}\sum_{\substack{\ell'=1 \\ \ell' \notin \{i,j\}}}^{n} z_{\ell'}\right)\right] \\
&= \sigma^2 E\left[\left(1 - \frac{1}{n}\right)\left(\frac{-1}{n}\right)z_i^2 + \left(1 - \frac{1}{n}\right)\left(\frac{-1}{n}\right)z_j^2 + \frac{1}{n^2}\sum_{\substack{\ell \notin \{i,j\} \\ \ell' \notin \{i,j\}}}^{n} z_\ell z_{\ell'} + \cancelto{0}{\Delta}\right] \\
&= -\sigma^2\left(1 - \frac{1}{n}\right)\left(\frac{1}{n}\right)E[z_i^2] - \sigma^2\left(1 - \frac{1}{n}\right)\left(\frac{1}{n}\right)E[z_j^2] + \sigma^2\frac{n-2}{n^2}E[z_i^2] \\
&= \sigma^2\left(-\frac{2}{n}\left(1 - \frac{1}{n}\right) + \frac{n-2}{n^2}\right) = \sigma^2\left(\frac{-2n + 2 + n - 2}{n^2}\right) = \frac{-\sigma^2}{n},
\end{aligned}
$$

where Δ is the set of cross terms which value is zero because all are in the form $E[z_i]E[z_{i'}] = 0$ $(i \neq i')$. ■

E.1.2 Two-Dimensional

If a $n \times m$ matrix $\boldsymbol{X}$ follows a two-dimensional degenerate normal with mean $\mu\underline{\mathbf{1}}_n$ and variance $\boldsymbol{\Sigma}$ then we write

$$\boldsymbol{X} \sim \text{DegenNormal}(n, m; \mu\underline{\mathbf{1}}_{nm}, \boldsymbol{\Sigma}),$$

where $\text{Var}(x_{ij}) = \sigma^2 \frac{(n-1)(m-1)}{nm}$, $\text{Cov}(x_{ij}, x_{ij'}) = \frac{-\sigma^2(n-1)}{nm}$ and $\text{Cov}(x_{ij}, x_{i'j}) = \frac{-\sigma^2(m-1)}{nm}$. In addition, the matrix $\boldsymbol{X}$ satisfies that each row and column have the same mean μ. This correspond to the $n + m - 1$ constraints

$$\begin{aligned} \frac{1}{m}\sum_{j=1}^{m} x_{ij} &= \mu \qquad \text{for } i = 1, \ldots, n, \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} &= \mu \qquad \text{for } j = 1, \ldots, m-1, \end{aligned} \tag{E.4}$$

which imply the additional relationship in the last column m: $\frac{1}{n}\sum_{i=1}^{n} x_{im} = \mu$. In the same manner as above for the one-dimensional case, we use two delta functions to apply these constraints and then the formulation of the density of this distribution is

$$\begin{aligned} f_{\text{Deg}\mathcal{N}}(\boldsymbol{X}|n, m, \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}(n-1)(m-1)} n^{\frac{1}{2}(m-1)} m^{\frac{1}{2}(n-1)} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\mu)^2\right) \\ \times \prod_{i=1}^{n} \delta\left(\sum_{j=1}^{m} x_{ij} - m\mu\right) \prod_{j=1}^{m-1} \delta\left(\sum_{i=1}^{n} x_{ij} - n\mu\right). \end{aligned} \tag{E.5}$$

Any individual component of $\boldsymbol{X}$ has a marginal normal distribution:

$$x_{ij} \sim \mathcal{N}\left(\mu, \frac{(n-1)(m-1)}{nm}\sigma^2\right) \qquad i = 1, \ldots, n \qquad j = 1, \ldots, m. \tag{E.6}$$

Proof eq. (E.6). We define nm independent and identically distributed random standard normal variables: $z_{ij} \sim \mathcal{N}(0,1)$ $(i = 1, \ldots, n$ and $i = j, \ldots, m)$.
Set $x_{ij} = \sigma(z_{ij} - \overline{z}_{i\cdot} - \overline{z}_{\cdot j} + \overline{z}_{\cdot\cdot}) + \mu$ as a linear combination of z_{ij} $(i = 1, \ldots, n$ and $i = j, \ldots, m)$, where $\overline{z}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} z_{ij}$, $\overline{z}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$, $\overline{z}_{\cdot\cdot} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}$, the hyperplane $\sum_{i=1}^{n} x_{ij}$ is

$$\begin{aligned}\sum_{i=1}^{n} x_{ij} &= \sigma\left(n\overline{z}_{\cdot j} - \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} z_{ij}\right) - \sum_{i=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\right) + \sum_{i=1}^{n}\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\right)\right) + n\mu \\ &= \sigma\left(n\overline{z}_{\cdot j} - \frac{1}{m}nm\overline{z}_{\cdot\cdot} - n\overline{z}_{\cdot j} + n\overline{z}_{\cdot\cdot}\right) + n\mu = n\mu,\end{aligned}$$

and the hyperplane $\sum_{j=1}^{m} x_{ij}$ is

$$\begin{aligned}\sum_{j=1}^{m} x_{ij} &= \sigma\left(m\overline{z}_{i\cdot} - m\overline{z}_{i\cdot} - \sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\right) + m\overline{z}_{\cdot\cdot}\right) + m\mu \\ &= \sigma\left(-\frac{1}{n}nm\overline{z}_{\cdot\cdot} + m\overline{z}_{\cdot\cdot}\right) + m\mu = m\mu.\end{aligned}$$

In addition,

$$\begin{aligned}x_{ij} &= \sigma\left(z_{ij} - \frac{1}{m}\sum_{j=1}^{m} z_{ij} - \frac{1}{n}\sum_{i=1}^{n} z_{ij} + \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\right) + \mu \\ &= \sigma\left(z_{ij} - \frac{1}{m}z_{ij} - \frac{1}{m}\sum_{\substack{\ell'=1\\ \ell'\neq j}}^{m} z_{i\ell'} - \frac{1}{n}z_{ij} - \frac{1}{n}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n} z_{\ell j} + \frac{1}{nm}z_{ij}\right) \\ &\quad + \sigma\left(\frac{1}{nm}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{\ell'=1\\ \ell'\neq j}}^{m} z_{\ell\ell'} + \frac{1}{nm}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n} z_{\ell j} + \frac{1}{nm}\sum_{\substack{\ell'=1\\ \ell'\neq j}}^{m} z_{i\ell'}\right) + \mu \\ &= \sigma\left(1 - \frac{1}{n} - \frac{1}{m} + \frac{1}{nm}\right) z_{ij} + \sigma\left(\frac{1}{nm} - \frac{1}{n}\right)\sum_{\substack{\ell=1\\ \ell\neq i}}^{n} z_{\ell j} \\ &\quad + \sigma\left(\frac{1}{nm} - \frac{1}{m}\right)\sum_{\substack{\ell'=1\\ \ell'\neq j}}^{m} z_{i\ell'} + \sigma\frac{1}{nm}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{\ell'=1\\ \ell'\neq j}}^{m} z_{\ell\ell'} + \mu.\end{aligned}$$

Therefore, x_{ij} is normally distributed with

$$E[x_{ij}] = \sigma\frac{(n-1)(m-1)}{nm}E[z_{ij}] + \sigma\left(\frac{1}{nm}-\frac{1}{n}\right)(n-1)E[z_{ij}] + \sigma\left(\frac{1}{nm}-\frac{1}{m}\right)(m-1)E[z_{ij}] + \sigma\frac{1}{nm}(n-1)(m-1)E[z_{ij}] + E[\mu] = \mu.$$

$$\begin{aligned} V[x_{ij}] &= \sigma^2\left(\frac{(n-1)(m-1)}{nm}\right)^2 V[z_{ij}] + \sigma^2\left(\frac{1}{nm}-\frac{1}{n}\right)^2(n-1)V[z_{ij}] \\ &\quad + \sigma^2\left(\frac{1}{nm}-\frac{1}{m}\right)^2(m-1)V[z_{ij}] + \sigma^2\left(\frac{1}{nm}\right)^2(n-1)(m-1)V[z_{ij}] + V[\mu] \\ &= \sigma^2\frac{(n-1)(m-1)}{nm}\left[\frac{(n-1)(m-1)}{nm} + \frac{(n-1)}{nm} + \frac{(m-1)}{nm} + \frac{1}{nm}\right] \\ &= \sigma^2\frac{(n-1)(m-1)}{nm}\left[\frac{1}{nm}\left(nm-n-m+1+n-1+m-1+1\right)\right] \\ &= \sigma^2\frac{(n-1)(m-1)}{nm}\left(\frac{1}{nm}nm\right) = \sigma^2\frac{(n-1)(m-1)}{nm}. \end{aligned}$$

■

Proof: The covariances $\mathbf{Cov}(x_{ij}, x_{ij'}) = \frac{-\sigma^2(n-1)}{nm}$ and $\mathbf{Cov}(x_{ij}, x_{i'j}) = \frac{-\sigma^2(m-1)}{nm}$.
We define nm independent and identically distributed random standard normal variables: $z_{ij} \sim \mathcal{N}(0,1)$ ($i = 1, \ldots, n$ and $i = j, \ldots, m$).
Set $x_{ij} = \sigma(z_{ij} - \overline{z}_{i\cdot} - \overline{z}_{\cdot j} + \overline{z}_{\cdot\cdot}) + \mu$ as a linear combination of z_{ij} ($i = 1, \ldots, n$ and $i = j, \ldots, m$), where $\overline{z}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} z_{ij}$, $\overline{z}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$, $\overline{z}_{\cdot\cdot} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}$, and $E[z_{ij}^2] = V[z_{ij}] + (E[z_{ij}])^2 = 1$ ($i = 1, \ldots, n$ and $i = j, \ldots, m$).
Thus, the covariance between any two different individuals components is formulated as:

$$
\begin{aligned}
\mathrm{Cov}(x_{ij}, x_{ij'}) &= E\left[(x_i - \overline{x})(x_j - \overline{x})\right] \\
&= E\left[(\sigma(z_{ij} - \overline{z}_{i\cdot} - \overline{z}_{\cdot j} + \overline{z}_{\cdot\cdot}) + \mu - \mu)\,(\sigma(z_{ij'} - \overline{z}_{i\cdot} - \overline{z}_{\cdot j'} + \overline{z}_{\cdot\cdot}) + \mu - \mu)\right] \\
&= \sigma^2 E\Bigg[\left(z_{ij} - \frac{1}{m}\sum_{j=1}^{m} z_{ij} - \frac{1}{n}\sum_{i=1}^{n} z_{ij} + \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\right) \\
&\qquad \left(z_{ij'} - \frac{1}{m}\sum_{j'=1}^{m} z_{ij'} - \frac{1}{n}\sum_{i=1}^{n} z_{ij'} + \frac{1}{nm}\sum_{i=1}^{n}\sum_{j'=1}^{m} z_{ij'}\right)\Bigg] \\
&= \sigma^2 E\Bigg[\Bigg(z_{ij} - \frac{1}{m}z_{ij} - \frac{1}{m}z_{ij'} - \frac{1}{m}\sum_{\substack{\ell=1 \\ \ell\neq\{j,j'\}}}^{m} z_{i\ell} - \frac{1}{n}z_{ij} - \frac{1}{n}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_{\ell j} \\
&\qquad + \frac{1}{nm}z_{ij} + \frac{1}{nm}z_{ij'} + \frac{1}{nm}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n}\sum_{\substack{\ell'=1 \\ \ell'\neq\{j,j'\}}}^{m} z_{\ell\ell'} + \frac{1}{nm}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_{\ell j} \\
&\qquad + \frac{1}{nm}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_{\ell j'} + \frac{1}{nm}\sum_{\substack{\ell=1 \\ \ell\neq\{j,j'\}}}^{m} z_{i\ell}\Bigg) \\
&\qquad \Bigg(z_{ij'} - \frac{1}{m}z_{ij} - \frac{1}{m}z_{ij'} - \frac{1}{m}\sum_{\substack{\ell=1 \\ \ell\neq\{j,j'\}}}^{m} z_{i\ell} - \frac{1}{n}z_{ij'} - \frac{1}{n}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_{\ell j'} \\
&\qquad + \frac{1}{nm}z_{ij} + \frac{1}{nm}z_{ij'} + \frac{1}{nm}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n}\sum_{\substack{\ell'=1 \\ \ell'\neq\{j,j'\}}}^{m} z_{\ell\ell'} + \frac{1}{nm}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_{\ell j} \\
&\qquad + \frac{1}{nm}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{n} z_{\ell j'} + \frac{1}{nm}\sum_{\substack{\ell=1 \\ \ell\neq\{j,j'\}}}^{m} z_{i\ell}\Bigg)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \sigma^2 E\Bigg[\Bigg(\frac{(n-1)(m-1)}{nm} z_{ij} + \left(\frac{1}{nm} - \frac{1}{m}\right) z_{ij'} + \left(\frac{1}{nm} - \frac{1}{m}\right) \sum_{\substack{\ell=1 \\ \ell \neq \{j,j'\}}}^{m} z_{i\ell} \\
&\qquad + \left(\frac{1}{nm} - \frac{1}{n}\right) \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} z_{\ell j} + \frac{1}{nm} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} z_{\ell j'} + \frac{1}{nm} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} \sum_{\substack{\ell'=1 \\ \ell' \neq \{j,j'\}}}^{m} z_{\ell\ell'} \Bigg) \\
&\qquad \Bigg(\frac{(n-1)(m-1)}{nm} z_{ij'} + \left(\frac{1}{nm} - \frac{1}{m}\right) z_{ij} + \left(\frac{1}{nm} - \frac{1}{m}\right) \sum_{\substack{\ell=1 \\ \ell \neq \{j,j'\}}}^{m} z_{i\ell} \\
&\qquad + \left(\frac{1}{nm} - \frac{1}{n}\right) \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} z_{\ell j'} + \frac{1}{nm} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} z_{\ell j} + \frac{1}{nm} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} \sum_{\substack{\ell'=1 \\ \ell' \neq \{j,j'\}}}^{m} z_{\ell\ell'} \Bigg)\Bigg] \\
&= \sigma^2 E\Bigg[\left(\frac{(n-1)(m-1)}{nm}\right)\left(\frac{1-n}{nm}\right) z_{ij}^2 + \left(\frac{(n-1)(m-1)}{nm}\right)\left(\frac{1-n}{nm}\right) z_{ij'}^2 \\
&\qquad + \left(\frac{1-n}{nm}\right)^2 \sum_{\substack{\ell \notin \{j,j'\} \\ \ell' \notin \{j,j'\}}}^{m} z_{i\ell} z_{i\ell'} + \left(\frac{1-m}{nm}\right)\left(\frac{1}{nm}\right) \sum_{\substack{\ell \neq i \\ \ell' \neq i'}}^{n} z_{\ell j} z_{\ell' j} \\
&\qquad + \left(\frac{1-m}{nm}\right)\left(\frac{1}{nm}\right) \sum_{\substack{\ell \neq i \\ \ell' \neq i'}}^{n} z_{\ell j'} z_{\ell' j'} + \left(\frac{1}{nm}\right)^2 \sum_{\substack{\ell \neq i \\ t \neq i'}}^{n} \sum_{\substack{\ell' \neq \{j,j'\} \\ t' \neq \{j,j'\}}}^{m} z_{\ell\ell'} z_{tt'} + \cancelto{0}{\Delta} \Bigg] \\
&= \sigma^2 \left(\frac{(n-1)(m-1)}{nm}\right)\left(\frac{1-n}{nm}\right) E[z_{ij}^2] + \sigma^2 \left(\frac{(n-1)(m-1)}{nm}\right)\left(\frac{1-n}{nm}\right) E[z_{ij'}^2] \\
&\qquad + \sigma^2 \left(\frac{1-n}{nm}\right)^2 (m-2) E[z_{ij}^2] + 2\sigma^2 \left(\frac{1-m}{nm}\right)\left(\frac{1}{nm}\right)(n-1) E[z_{ij}^2] \\
&\qquad + \sigma^2 \left(\frac{1}{nm}\right)^2 (n-1)(m-2) E[z_{ij}^2] \\
&= \frac{-\sigma^2(n-1)}{nm}\left(\frac{1}{nm}\left(2(n-1)(m-1) + (1-n)(m-2) - 2(1-m) - (m-2)\right)\right) \\
&= \frac{-\sigma^2(n-1)}{nm}\left(\frac{1}{nm} nm\right) = \frac{-\sigma^2(n-1)}{nm},
\end{aligned}
$$

where Δ is the set of cross terms which value is zero because all are in the form $E[z_{ij}]E[z_{ij'}] = 0$ $(j \neq j')$.

The proof of $\text{Cov}(x_{ij}, x_{i'j}) = \frac{-\sigma^2(m-1)}{nm}$ has the same steps exchanging rows (n) by columns (m). ■

Appendix F

Convergence Diagnostics for MCMC

The purpose of convergence diagnostics is to determine when it is reasonable to believe that the samples generated by MCMC samplers are representative of the underlying target probability distribution (i.e. the posterior distribution). The interest relies on how well the chain is mixing over the parameter space in order to obtain reliable parameter estimates. In this thesis, we use four of the most common convergence diagnostic tests in the literature to assess whether a chain has converged to the stationary distribution: Geweke time series diagnostic (Geweke (1992), Section F.1), Gelman and Rubin's multiple sequence diagnostic (Gelman and Rubin (1992); Brooks (1998), Section F.2), Heidelberger and Welch diagnostic (Heidelberger and Welch (1983), Section F.3), and effective sample size (ESS) (Kass et al. (1998), Section F.4).

The following sections describe these methods and their technical details are outlined.

F.1 Geweke Time Series Diagnostic

The Geweke time series diagnostic proposed in Geweke (1992) is based on the comparison of the means of parameters' posterior distributions from two non-overlapping portions of a single MCMC chain by using a test for equality of the means. The portions are usually the first 10% of draws and the last 50% of draws from the complete Markov chain. If a model has converged, then the mean from the first portion of the chain will be approximately equal to the mean from the second portion of the chain.

The test statistic is a standard Z-score calculated by the difference between the two sample means divided by their estimated standard error. The standard error is estimated from the spectral density at zero and so takes into account any autocorrelation. This test is calculated assuming that the 2 portions of the chain are asymptotically independent (non-overlapping portions) and is asymptotically distributed as a standard normal distribution.

We suppose Ω is the parameter vector of interest. For simplicity of notation, Ω is assumed to be one-dimensional in this section. Define $\{\Omega^t\}$, where $t = 1, \ldots, T$, to be a single MCMC output of length T. The procedure to calculate the Geweke time series diagnostic is as follows:

1. Preselect two non-overlapping portions from the Markov chain: $\{\Omega_1^{t_1} : t_1 = 1, \ldots, n_1\}$ of length n_1 and $\{\Omega_2^{t_2} : t_2 = 1, \ldots, n_2\}$ of length n_2 .

2. Calculate the Geweke's standard Z-score:

$$G = \frac{\overline{\Omega}_1 - \overline{\Omega}_2}{\sqrt{\frac{s_1(0)}{n_1} + \frac{s_2(0)}{n_2}}} \sim N(0, 1),$$

   where

$$\overline{\Omega}_j = \frac{\sum_{t_j=1}^{n_j} \Omega_j^{t_j}}{n_j}, \tag{F.1}$$

   is the mean of the portion j ($j = 1, 2$), i.e. $\{\Omega_j^{t_j} : t_j = 1, \ldots, n_j\}$ and $s_j(0)$ ($j = 1, 2$) is the estimated standard error by using the spectral density at zero (i.e. it does not take into account any autocorrelation) defined for the time series confined within the corresponding portion.

3. Interpret Geweke's diagnostic G as

$$|G| = \begin{cases} \leq 2 & \text{The chain has converged} \\ \\ > 2 & \text{The chain has not converged} \end{cases}.$$

F.2 Gelman and Rubin's Multiple Sequence Diagnostic

The Gelman and Rubin's multiple sequence diagnostic proposed in Gelman and Rubin (1992) and is based on the comparison of a set of chains drawn with different starting points which are overdispersed relative to the target distribution. The comparison uses within and between chain variances for each parameter in order to test whether the set of Markov chains are overlapping. The criterion to assess whether the Markov chain converges is contrasting the variance within and the variance between chains. We assume Ω is the parameter vector to estimate and a set of m Markov chains each of length $2n$ are available for Ω. The procedure to calculate the Gelman and Rubin's multiple sequence diagnostic is as follows:

1. Run $m \geq 2$ chains of lengths $2n$ each from over-dispersed starting values. In that manner, the set $\{\Omega_j^t : t = 1, \ldots, 2n\}$ is generated for $j = 1, \ldots, m$.

2. Discard the first half of draws (n) in each chain as burn in.

3. Calculate the *within-chain variance* W as

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2,$$

   where s_j^2 is the variance of the j^{th} chain:

$$s_j^2 = \frac{1}{n-1} \sum_{t=1}^{n} \left(\Omega_j^t - \overline{\Omega}_j\right)^2.$$

   The within-chain variance W could underestimate the true variance of the target distribution of Ω because all m chains may not cover the full support of the stationary distribution.

4. Calculate the *between-chain variance* B as

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\overline{\Omega}_j - \overline{\overline{\Omega}}\right)^2,$$

where $\overline{\overline{\Omega}}$ is the mean of the means of the chains defined as

$$\overline{\overline{\Omega}} = \frac{1}{m}\sum_{j=1}^{m}\overline{\Omega}_j.$$

5. Calculate the estimated variance of the parameter vector Ω as

$$\widehat{\text{Var}(\Omega)} = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B,$$

which is the weighted sum of the within-chain W and between-chain variance B.

6. Calculate the potential scale reduction factor (PSRF) as

$$\text{PSRF} = \sqrt{\frac{\widehat{\text{Var}(\Omega)}}{W}}.$$

7. Evaluate PSRF as

$$\text{PSRF} = \begin{cases} \leq 1.2 & \text{Convergence has been achieved} \\ \\ > 1.2 & \text{No convergence. Longer chain is required} \end{cases}.$$

When the convergence is reached ($\text{PSRF} \leq 1.2$) means that the mn draws from the m chains derive from the same stationary distribution (overlapped chains) and therefore we can pool them to produce a set of draws from the target distribution.

F.3 Heidelberger and Welch Diagnostic

The Heidelberger and Welch diagnostic (Heidelberger and Welch, 1983) is based on the Cramér-von Mises test statistic to evaluate the null hypothesis that the values drawn from a Markov chain come from a stationary distribution. This diagnostic consists of two tests: a stationary test and a halfwidth test. The stationary test is an iterative test which is successively applied, firstly to the whole chain, then after removing progressively portions of 10% of the draws from the

first draws of the chain. This procedure is repeated until either the null hypothesis is accepted, or 50% of the chain has been discarded. Moreover, the halfwidth test validates the results obtained in the stationary test when the null hypothesis is accepted and is based on the computation of a (1-α)% credible interval for the sampled mean (α is the type I error of the hypothesis test).

The procedure to calculate the Heidelberger and Welch diagnostic is as follows:

- *Stationary test*:

  1. Generate a chain of n draws and specify a type I error level α for the test (the most common value is $\alpha = 0.05$).
  2. Calculate the test statistic on the whole Markov chain. This test is based on the Cramer-von Mises statistic; that is $\int_0^1 B_n(\ell)^2 \mathrm{d}\ell$ where $B_n(\ell)$ ($\ell = \frac{1}{n}, \frac{2}{n}, \ldots, 1$) is a sequence formulated as:

  $$B_n(\ell) = \frac{(S_{[n\ell]} - [n\ell]\overline{\Omega})}{\sqrt{ns(0)}},$$

  where $S_0 = 0$, $S_n = \sum_{t=1}^{n} \Omega^t$, and $[\,]$ is the rounding operator. The function $s(0)$ is the symmetric spectral density function at zero and $\overline{\Omega}$ is the mean for the whole MCMC as shown in eq. (F.1) in Section F.1.
  3. Decide whether to accept or reject the null hypothesis H_0 : the chain has reached stationarity, by using the previous test statistic. The common significance level value of this test is $\alpha = 0.05$.
  4. If the null hypothesis is rejected, then remove the first 10% draws from the Markov chain and return to step 2. Repeat this iterative procedure until either the 50% of the chain is removed or the null hypothesis is accepted.
  5. If the null hypothesis is accepted, then we perform the halfwidth test. Otherwise, we need to increase the number of draws n and return to step 1. In their paper, Heidelberger and Welch suggested increasing the run length by a factor higher than 1.5 each time so that there is a reasonably large proportion of new draws and we avoid problems caused by sequential testing due to repeating the test too frequently on the same data.

- *Halfwidth test*:

  1. Take the part of the Markov chain which remains at the end of the stationary test.
  2. Calculate half of the width HW of the $100(1-\alpha)\%$ (commonly $\alpha = 0.05$) credible interval around the sample mean $\overline{\Omega}$:

$$\mathrm{HW} = \frac{\mathrm{halfwidth}}{\mathrm{mean}} = \frac{\mathrm{t}_{\mathrm{n},\alpha/2}\frac{\mathrm{sd}(\overline{\Omega})}{\sqrt{\mathrm{n}}}}{\overline{\Omega}}.$$

  3. Choose the required relative error ε (the most common value is $\varepsilon = 0.1$, i.e. 10%).
  4. Evaluate the following

$$\mathrm{HW} = \begin{cases} < \varepsilon & \text{Convergence has been achieved} \\ \\ \geq \varepsilon & \text{No convergence. Longer chain is required} \end{cases}.$$

F.4 Effective Sample Size

The effective Sample Size (ESS) diagnostic was recommended by Radford Neal in the panel discussion of Kass et al. (1998). This measure is based on the principle that the higher the autocorrelation in the MCMC samples, the lower the information in the posterior distributions. In other words, the autocorrelation reduces the effective sample size of representing the posterior distribution.

The ESS is formulated as follows:

$$\mathrm{ESS} = \frac{T}{1 + 2\sum_{k=1}^{\infty} \rho_k(\omega)},$$

where T is the number of posterior samples, ω is the parameter vector of marginal posterior samples, and ρ_k is the autocorrelation function for ω at lag k.

The ESS is a quantity that estimates the number of independent draws in the chain. Thus, a smaller ESS is due to correlated draws in the chain. Stopping the MCMC updates is not recommended until the ESS is greater than some threshold value (the most common threshold is ESS ≥ 200).

Appendix G

A Relabelling Algorithm for Mixture Models

The relabelling procedure we implemented to overcome the label switching problem for our finite mixture model approach closely follows Stephens (2000a, Section 4.1). We develop here the row clustering version. The column clustering version is the same except exchanging rows for columns. The application of this procedure for the biclustering approach case consists of two steps: First relabelling in one dimension (e.g. rows) and, once the label switching problem is solved for that dimension, running it for the other dimension (e.g. columns).

We define R as the number of clusters, $\boldsymbol{Y} = (y_{ij})$ as an ordinal $n \times m$ data matrix, q as the number of ordinal categories, T as the length of the chain sample (after burn-in period), $\Omega^{(t)}$ as the parameter vector at iteration t, $\theta^{(t)}$ as the set of parameters involved in the label switching problem at iteration t, $\eta^{(t)}$ as the set of parameters from $\Omega^{(t)}$ not included in $\theta^{(t)}$, and $\nu_t(\theta)$ is the permutation of the parameter vector θ at iteration t. The algorithm can be summarised in the following iterative steps:

1. Specify an arbitrary initial permutation values for $\nu_1, \ldots, \nu_T$ (Stephens (2000a) suggested the identity permutation).

2. Compute the $n \times R$ matrix $\widehat{Z} = (\widehat{z}_{ir})$ where

$$\widehat{z}_{ir} = \frac{1}{T} \sum_{t=1}^{T} p_{ir}[\nu_t(\theta^{(t)})] \tag{G.1}$$

and

$$p_{ir}[\nu_t(\theta^{(t)})] = \frac{\widehat{\pi}_r^{(t)} \prod_{j=1}^{m} \prod_{k=1}^{q} \left(P[y_{ij} = k | i \in r, \eta^{(t)}, \nu_t(\theta^{(t)})]\right)^{I(y_{ij}=k)}}{\sum_{\ell=1}^{R} \left\{ \widehat{\pi}_\ell^{(t)} \prod_{j=1}^{m} \prod_{k=1}^{q} \left(P[y_{ij} = k | i \in r, \eta^{(t)}, \nu_t(\theta^{(t)})]\right)^{I(y_{ij}=k)} \right\}},$$

is the posterior probability that subject i is classified in cluster r (as defined in eq. in (2.18) in Section 2.5.1), once we have observed the data $\{y_{ij}\}$ and given the permutation values $\nu_t(\theta^{(t)})$ at iteration t.

3. For $t = 1, 2, \ldots, T$ choose ν_t to minimize

$$\sum_{i=1}^{n} \sum_{r=1}^{R} p_{ir}[\nu_t(\theta^{(t)})] \log \left(\frac{p_{ir}[\nu_t(\theta^{(t)})]}{\widehat{z}_{ir}} \right). \tag{G.2}$$

   The most easy way to achieve this step is by exploring all $R!$ possible rearrangements for each ν_t, and select the one with the lowest value in eq. (G.2). The matrix Z is recalculated in each possible rearrangement, as its values depend on $\nu_t(\theta^{(t)})$ (see eq. (G.1)).

4. Test whether a fixed point is reached. Otherwise, return to step 2.
   A fixed point is when $\nu_t(\theta^{(t)}) = \nu_{(t-1)}(\theta^{(t-1)})$, i.e. there is no change in the values of the permutation of θ at iteration t and $t-1$.

Appendix H

Metropolis-Hastings Sampler

Section H.1 outlines a simulation study we have carried out to test the reliability of the M-H sampler for our one-dimensional clustering approach. Section H.2 depicts the results of the simulation study for several scenarios.

H.1 Simulation Study. Outline

The simulation study procedure for the row clustering model by using a M-H sampler is outlined in this Section. The simulation study outline for the column clustering version is basically the same to the row clustering version just replacing parameters related to rows for their equivalent to columns.

Step 1. Model specification

Select the model, w, from a set of models $w = 1, \ldots, W$.
There are in total $W = (1 \times 3) + 2 = 5$ possible models:

- Set the number of response categories: $q = 4$ in all cases (1 option).
  This fixes $\{\mu_1, \ldots, \mu_q\}$ (with $\mu_1 = 0$).
- Select $R \in \{2, 3, 4\}$ (3 options + 2 special cases).
  This fixes $\{\alpha_1, \ldots, \alpha_R\}$ (with $\sum_{r=1}^{R} \alpha_r = 0$).
  If $R = 3$ there are three possible scenarios to select. Two of them are special scenarios:
  - two adjacent response categories having equal values, and

- one component with a very small prior mixing probability.

Each scenario fixes:

- Number of columns $m \in \{3, 5\}$ ($m = 5$ only in the 2 special cases) This fixes $\{\beta_1, \ldots, \beta_m\}$ (with $\sum_{j=1}^{m} \beta_j = 0$).
- Prior mixing probabilities $\pi_1, \ldots, \pi_R$ (with $\sum_{r=1}^{R} \pi_r = 1$).
- The ordinal response cut levels $\phi_1 \leq \phi_2 \leq \ldots \leq \phi_q$ (with $\phi_1 = 0$ and $\phi_q = 1$).

At the end of this step we know, for the chosen model w:

- The number of row groups R^w.
- The number of response categories q^w.
- The number of columns m^w.
- The total number of free parameters K^w.
- The parameter values:

$$\{\alpha_1^w, \ldots, \alpha_R^w\}, \{\beta_1^w, \ldots, \beta_m^w\}, \{\pi_1^w, \ldots, \pi_R^w\}, \{\mu_1^w, \ldots, \mu_q^w\}, \{\phi_1^w, \ldots, \phi_q^w\}$$

and as a consequence we can calculate the values of the linear predictors

$$\eta_{krj}^w = \mu_k^w + \phi_k^w \left(\alpha_r^w + \beta_j^w\right)$$

for $k \in \{1, \ldots, q^w\}$, $r \in \{1, \ldots, R^w\}$ and $j \in \{1, \ldots, m^w\}$.

Step 2. Simulator specification

Set the parameters for the simulator specifying:

- The number of replicates to run (i.e. distinct datasets): $H = 100$.
- The number of chains in each replicate: $S = 3$.

Step 3. Markov Chain specification

Set the chain parameters specifying:

- The number of iterations in the burn-in period: nburn=20000.
- The number of iterations to store: nstore=20000.
- The thinning rate: nthin=5.

Step 4. MH parameters and hyperparameters specification

Set the hyperparameter values specifying:

- Shape and scale parameters to specify an Inverse Gamma distribution which is the prior for the variance parameter from Normal distributions for each of
  - the cut point parameters $\{\mu_k\}$: $\nu_\mu = 3$, $\delta_\mu = 40$,
  - the row cluster parameters $\{\alpha_r\}$: $\nu_\alpha = 3$, $\delta_\alpha = 40$, and
  - the column parameters $\{\beta_j\}$: $\nu_\beta = 3$, $\delta_\beta = 40$.
- Parameter vector for a Dirichlet distribution for each of
  - the score parameters $\{\phi_k\}$: $\boldsymbol{\lambda}_\phi = \mathbf{1}$, and
  - the prior mixing probabilities $\{\pi_r\}$: $\boldsymbol{\lambda}_\pi = \mathbf{1}$.

Step 5. Proposal parameters specification

Set the parameter values for all the proposal distributions $q(\cdot|\cdot)$ to:

- an update of the cut point parameters $\{\mu_k\}$: $\sigma^2_{\mu_p} = 0.3$,
- an update of the row cluster parameters $\{\alpha_r\}$: $\sigma^2_{\alpha_p} = 0.3$,
- an update of the column parameters $\{\beta_j\}$: $\sigma^2_{\beta_p} = 0.3$, and
- an update of the row group membership probability parameters $\{\pi_r\}$: $\sigma^2_{\pi_p} = 0.09$.

Step 6. Generate replicate datasets

For each replicate $h \in \{1, \ldots, H\}$ and each chain $s \in \{1, \ldots, S\}$:

- For each row $i = 1, \ldots, n$, generate row membership as an indicator vector

$$\mathbf{z}_i^{hs} = \left(Z_{i1}^{hs}, ..., Z_{iR}^{hs}\right) \sim \text{Multinomial}\left(1; \{\pi_r\}\right).$$

- For each column $j = 1, \ldots, m$ within each row $i = 1, \ldots, n$, generate the response ordinal variable

$$y_{ij}^{hs} | \mathbf{z}_i^{hs} = \boldsymbol{\delta}_r \sim \text{Stereotype}\left(\{\eta_{krj}\}_{k=1}^{q}\right).$$

  Here $\boldsymbol{\delta}_r$ is an indicator vector of length R, with 1 at location r and zero elsewhere. This implies that

$$\log\left(\frac{\mathrm{P}\left[y_{ij}^{hs} = k \mid \mathbf{z}_i^{hs} = \boldsymbol{\delta}_r\right]}{\mathrm{P}\left[y_{ij}^{hs} = 1 \mid \mathbf{z}_i^{hs} = \boldsymbol{\delta}_r\right]}\right) = \eta_{krj} = \mu_k + \phi_k(\alpha_r + \beta_j).$$

Step 7. Fit model. Run the Metropolis-Hastings sampler

- We run the Metropolis-Hastings sampler for the dataset hs.
  On iteration t ($t = 1, \ldots, \text{nstore}$) we obtain the estimated parameter vector $\widehat{\Omega}_{(t)}^{hs}$ for the dataset hs consisting of parameters:

$$\{\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{R-1}\}, \{\widehat{\beta}_1, \ldots, \widehat{\beta}_{m-1}\}, \{\widehat{\pi}_1, \ldots, \widehat{\pi}_{R-1}\}, \{\widehat{\mu}_2, \ldots, \widehat{\mu}_q\}, \text{ and } \{\widehat{\phi}_2, \ldots, \widehat{\phi}_{q-1}\}.$$

- Return the values $\left\{\widehat{\Omega}_{(1)}^{hs}, \widehat{\Omega}_{(2)}^{hs}, \ldots, \widehat{\Omega}_{(\text{nstore})}^{hs}\right\}$ for $h = 1, \ldots, H$ and $s = 1, \ldots, S$.
- Test whether the convergence has been achieved. If not, increase nstore and run the Metropolis-Hasting sampler again.
- Test whether the label-switching problem is observed in the posterior distributions of $\{\alpha_r\}$ and $\{\pi_r\}$. If so, perform the procedure described in Section 7.2.5 and Appendix G.

Step 8. Summarising results

- For each replicate, merge the s chains into one chain.

- Summarise the results by computing the mean, median, standard deviation, time series standard error and highest posterior density interval (HPD) for each element of the parameter vectors $\{\widehat{\Omega}^h_{(1)}, \ldots, \widehat{\Omega}^h_{(\text{nstore})}\}$.

- Report the marginal posterior distribution and trace plot for each element of the parameter vectors $\{\widehat{\Omega}^h_{(1)}, \ldots, \widehat{\Omega}^h_{(\text{nstore})}\}$.
  One way to report this is sampling 10000 iterations from all the pooled replicates H in order to reduce the storage and make the depiction easier.

- Report the following convergence diagnostics,

  - PSRF from Gelman and Rubin's multiple sequence diagnostic.
  - Gelman-Rubin-Brooks plots showing the evolution of Gelman and Rubin's shrinkage factor.
  - Marginal posterior distributions and trace plots overlapping the S chains.

H.2 Simulation Study. Results

Figures H.1-H.10 show the marginal posterior distribution and trace plot for all the parameters for $R = 2, 3, 4$ number of row clusters respectively and the two special cases. The expected values of the posterior distribution are very close to the true values (green vertical lines) and the 95% HPD credible interval includes the true parameter values in all the cases. The trace plots on Figures H.2, H.4, and H.6 show a good mixing on all the parameters. Additionally, Figure H.11 shows the Gelman and Rubin diagnostic plots to assess convergence over the $S = 3$ chains and for one particular replicate. The evolution of the shrinkage factor as the number of iterations increases shows good convergence in the MCMC output from the Metropolis-Hasting sampler.

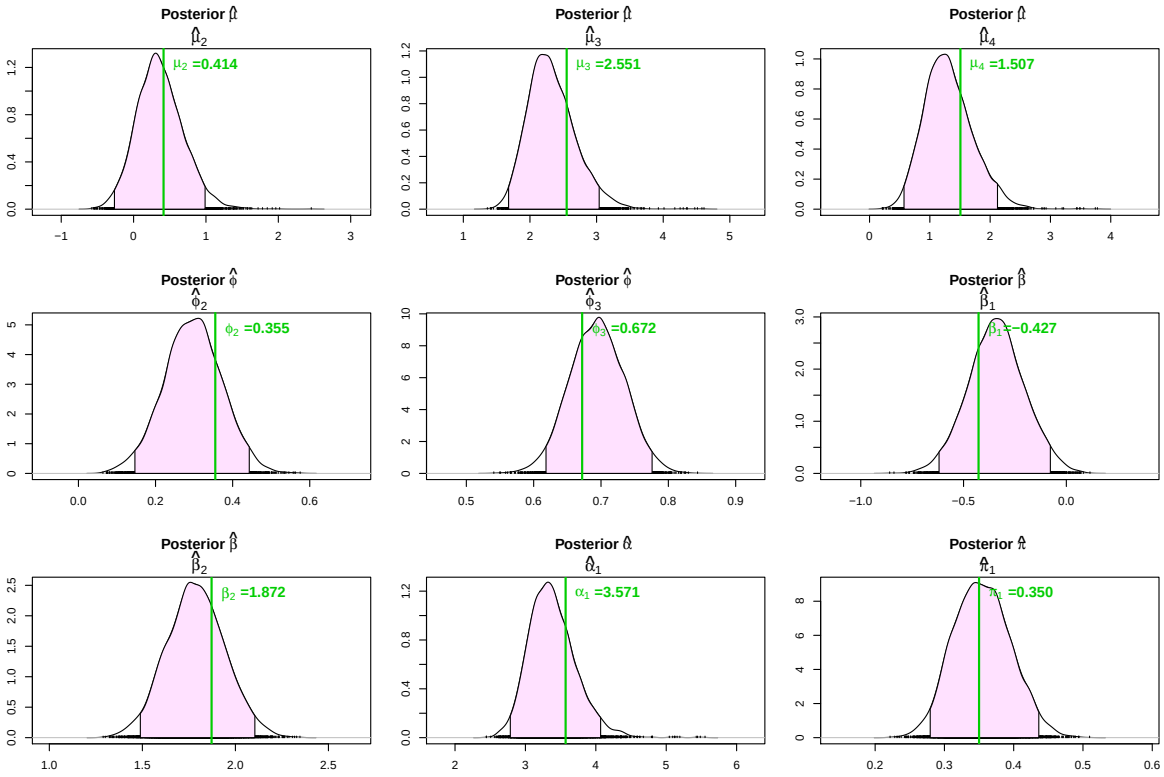

Figure H.1: ***MH simulation study***: *Density plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. *The sample size is* $n = 500$, *the number of categories is* $q = 4$, *the number of columns is* $m = 3$ *and the number of row clusters is* $R = 2$. *The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 3$*). The green vertical lines are the true parameter values and 95% HPD credible intervals are shown with shading area.*

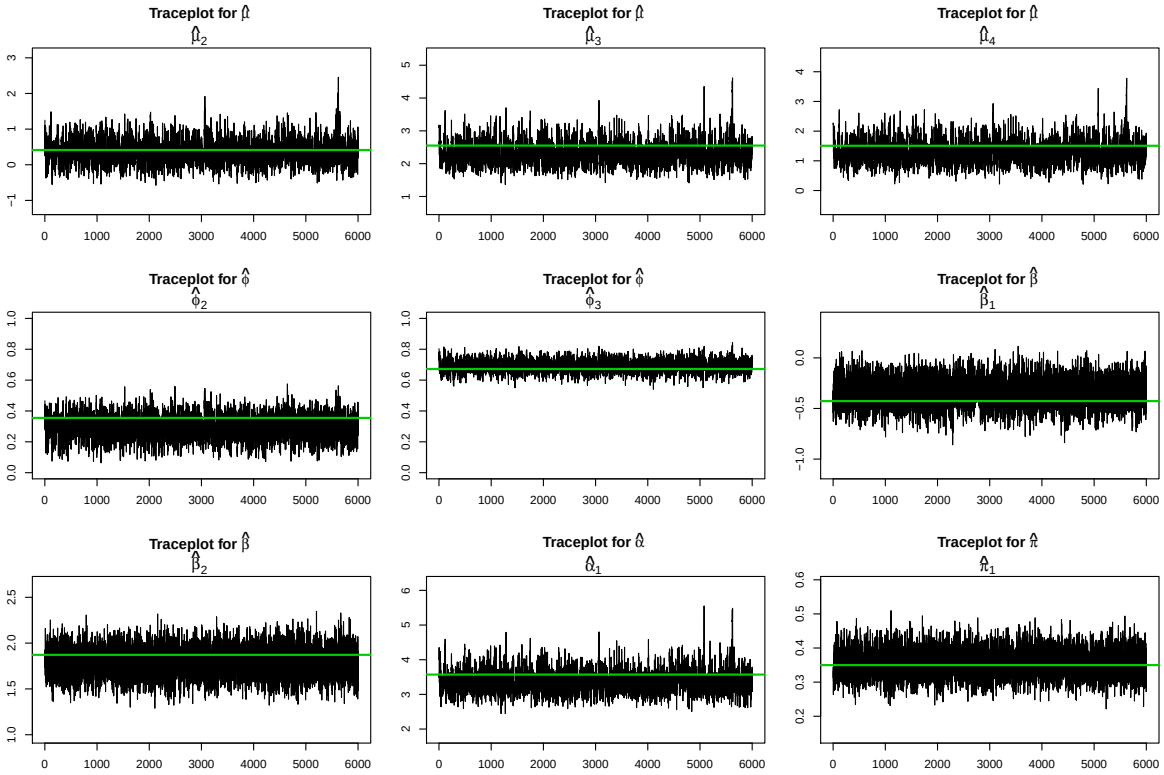

Figure H.2: ***MH simulation study***: *Trace plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. *The sample size is* $n = 500$, *the number of categories is* $q = 4$, *the number of columns is* $m = 3$ *and the number of row clusters is* $R = 2$. *The plots depict the results of the Metropolis-Hastings sampler for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 3$*). The green horizontal lines are the true parameter values.*

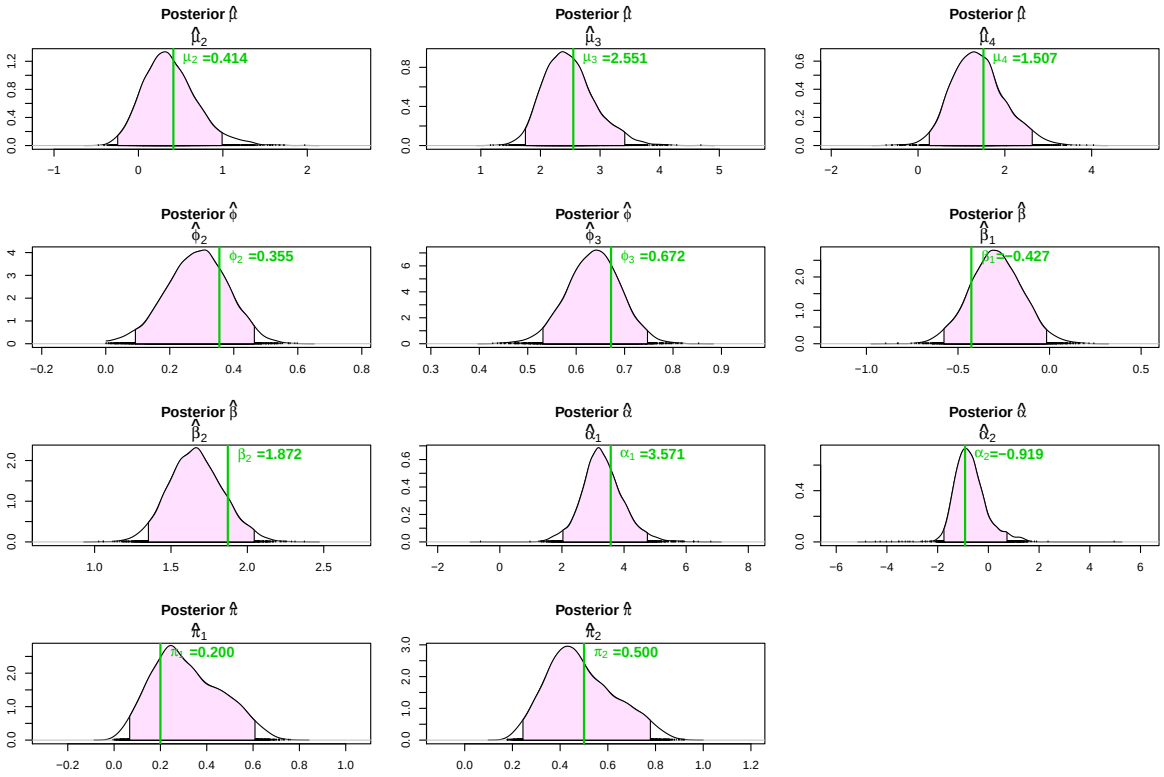

Figure H.3: ***MH simulation study***: *Density plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. *The sample size is* $n = 500$, *the number of categories is* $q = 4$, *the number of columns is* $m = 3$ *and the number of row clusters is* $R = 3$. *The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 3$*). The green vertical lines are the true parameter values and 95% HPD credible intervals are shown with shading area.*

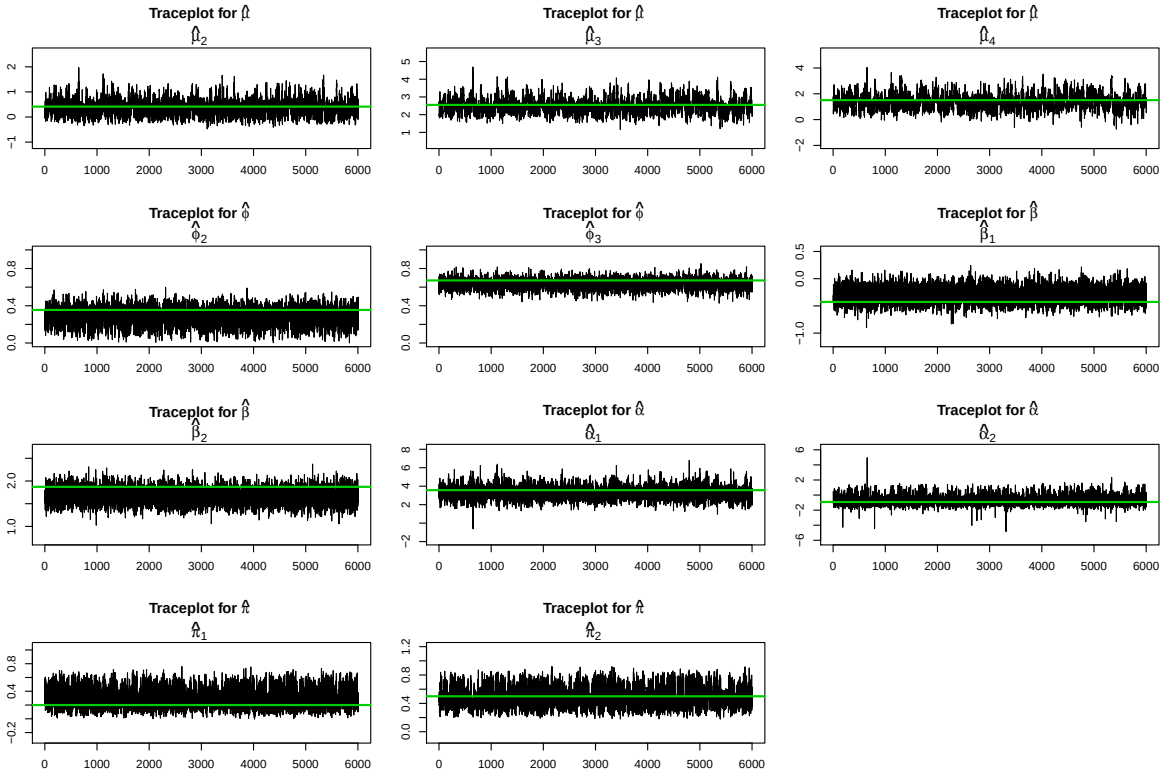

Figure H.4: ***MH simulation study****: Trace plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$*. The sample size is* $n = 500$*, the number of categories is* $q = 4$*, the number of columns is* $m = 3$ *and the number of row clusters is* $R = 3$*. The plots depict the results of the Metropolis-Hastings sampler for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 3$*). The green horizontal lines are the true parameter values.*

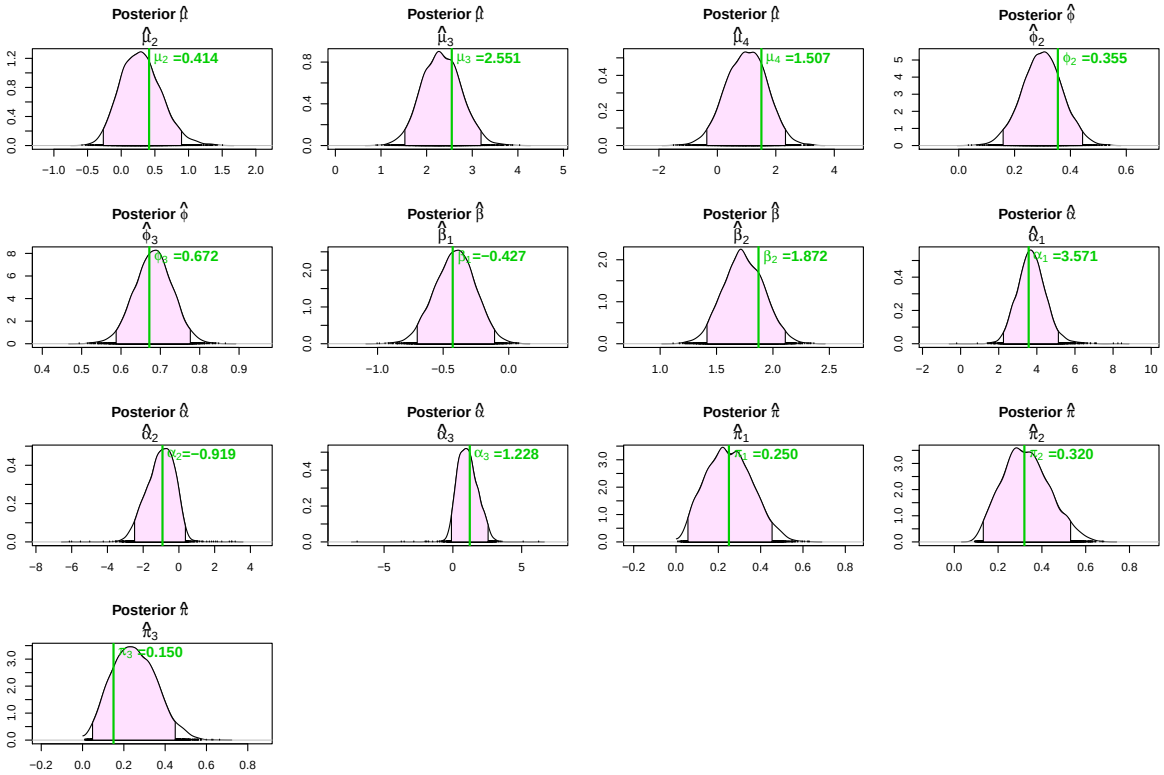

Figure H.5: ***MH simulation study***: *Density plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. *The sample size is* $n = 500$, *the number of categories is* $q = 4$, *the number of columns is* $m = 3$ *and the number of row clusters is* $R = 4$. *The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 3$*). The green vertical lines are the true parameter values and 95% HPD credible intervals are shown with shading area.*

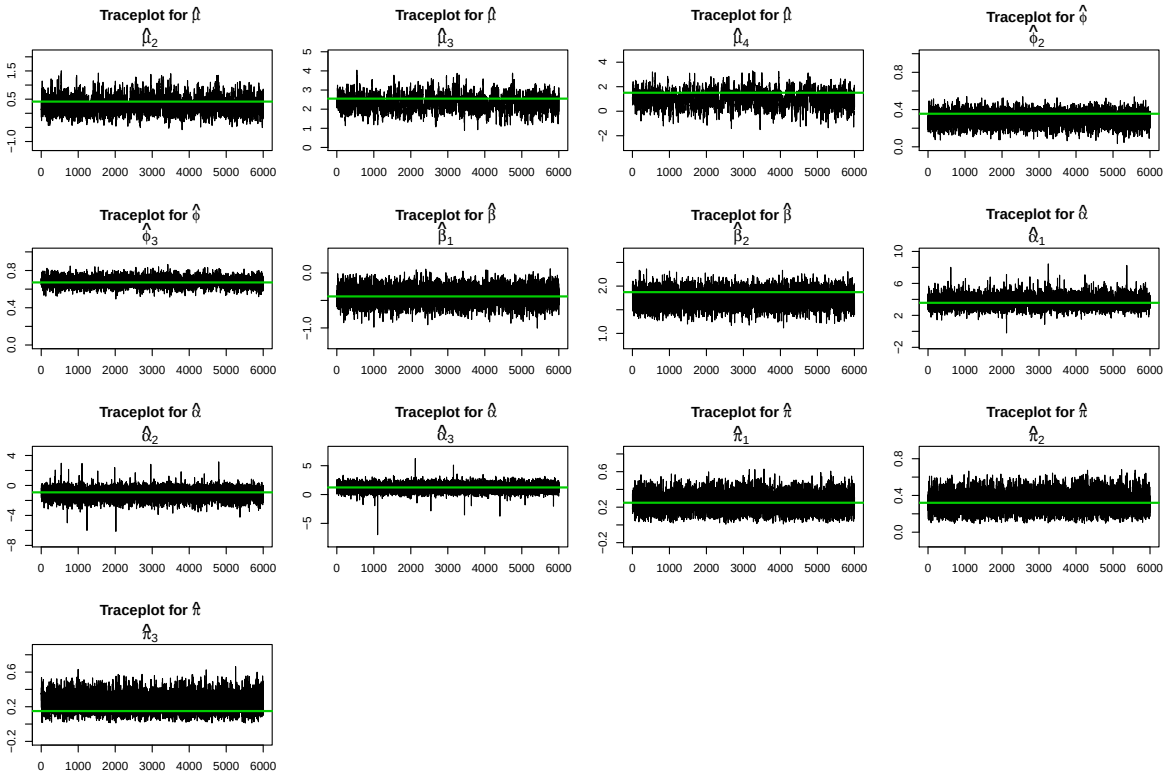

Figure H.6: ***MH simulation study****: Trace plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$*. The sample size is* $n = 500$*, the number of categories is* $q = 4$*, the number of columns is* $m = 3$ *and the number of row clusters is* $R = 4$*. The plots depict the results of the Metropolis-Hastings sampler for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 3$*). The green horizontal lines are the true parameter values.*

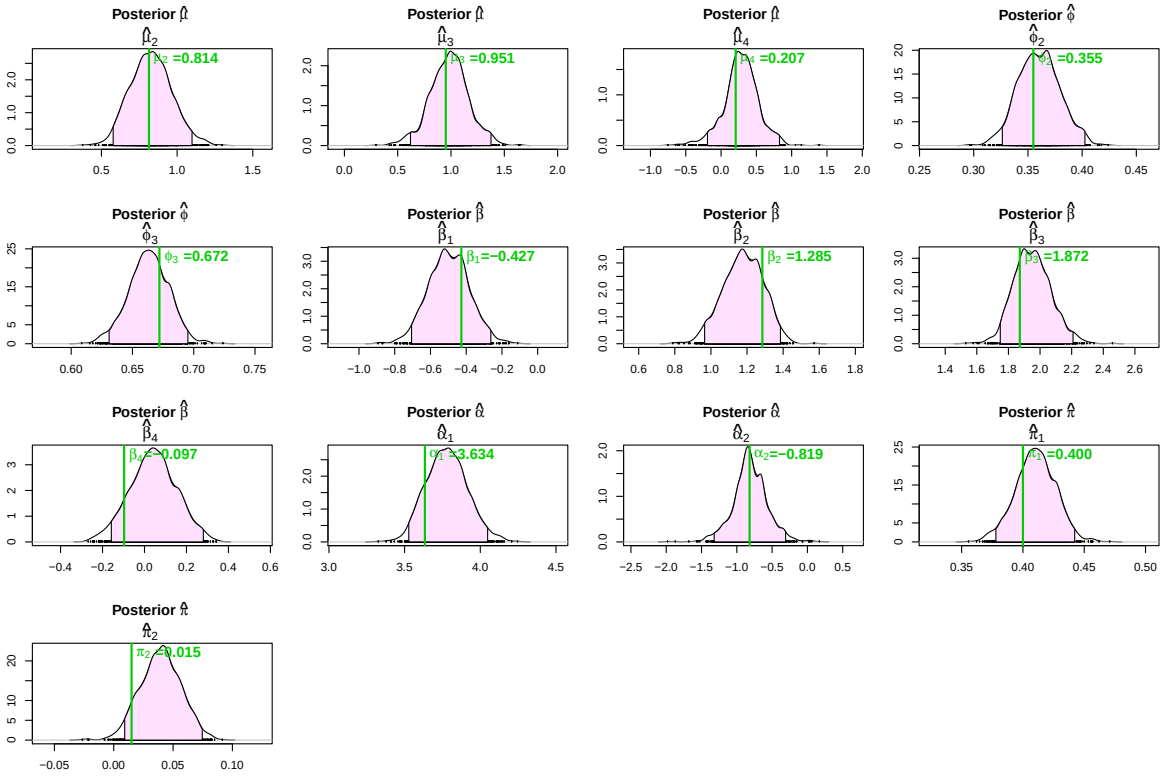

Figure H.7: ***MH simulation study***: *Density plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$ *when the parameter* π_2 *takes a very small value. The sample size is* $n = 500$*, the number of categories is* $q = 4$*, the number of columns is* $m = 5$ *and the number of row clusters is* $R = 3$*. The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates* ($H = 100$) *and chains* ($S = 3$)*. The green vertical lines are the true parameter values and 95% HPD credible intervals are shown with shading area.*

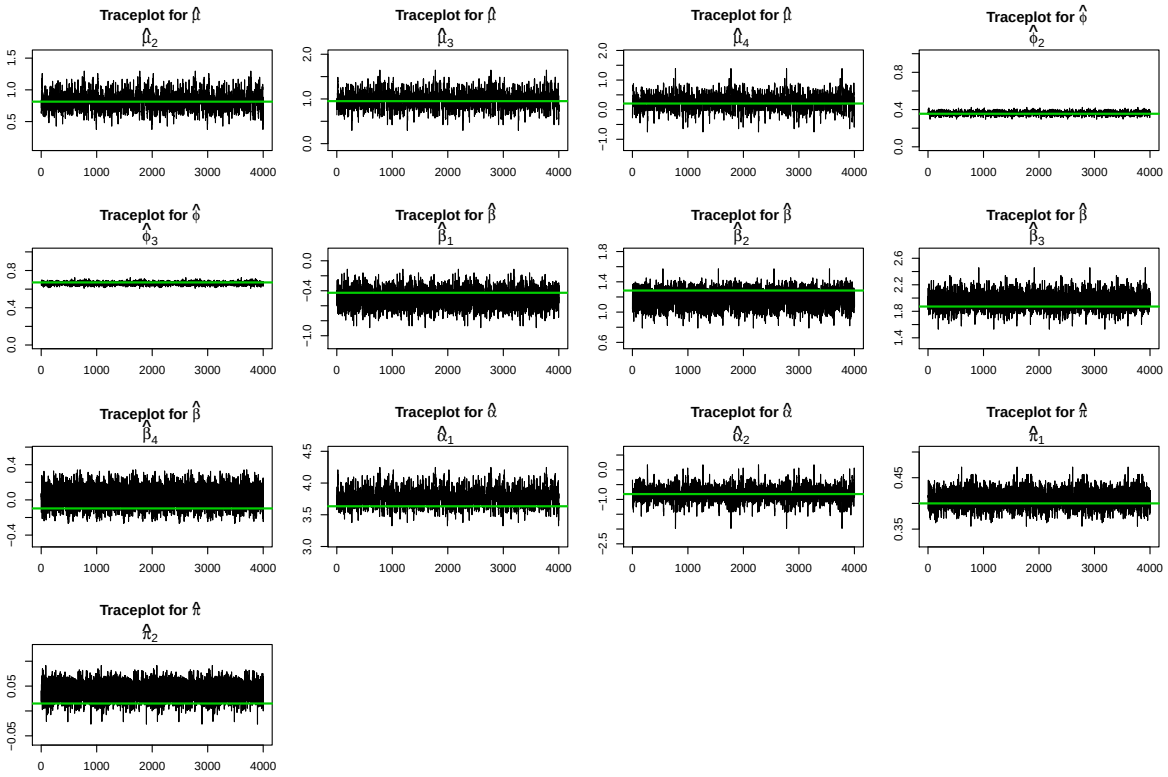

Figure H.8: ***MH simulation study****: Trace plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$ *when the parameter* π_2 *takes a very small value. The sample size is* $n = 500$*, the number of categories is* $q = 4$*, the number of columns is* $m = 5$ *and the number of row clusters is* $R = 4$*. The plots depict the results of the Metropolis-Hastings sampler for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 3$*). The green horizontal lines are the true parameter values.*

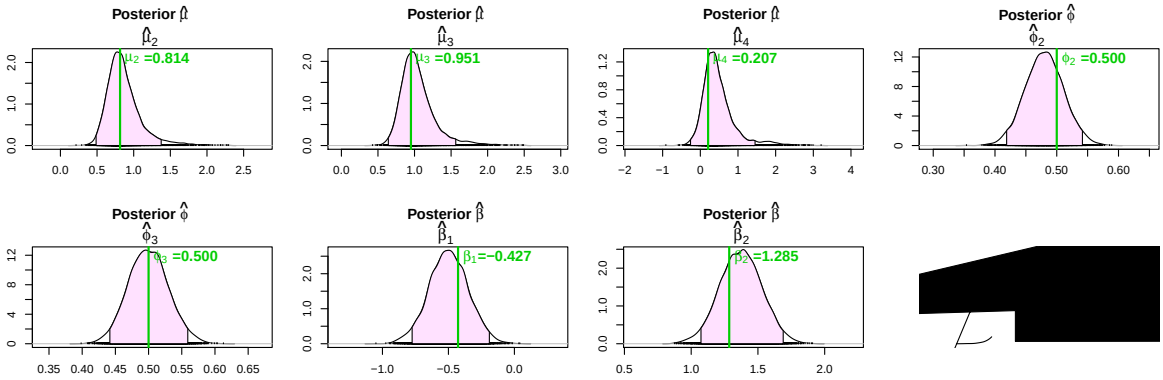

Figure H.9: ***MH simulation study***: *Density plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k + \phi_k(\alpha_r + \beta_j)}$ *when the adjacent parameters* $\phi_2 = \phi_3$*. The sample size is* $n = 500$*, the number of categories is* $q = 4$*, the number of columns is* $m = 5$ *and the number of row clusters is* $R = 3$*. The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates* ($H = 100$) *and chains* ($S = 3$)*. The green vertical lines are the true parameter values and 95% HPD credible intervals are shown with shading area.*

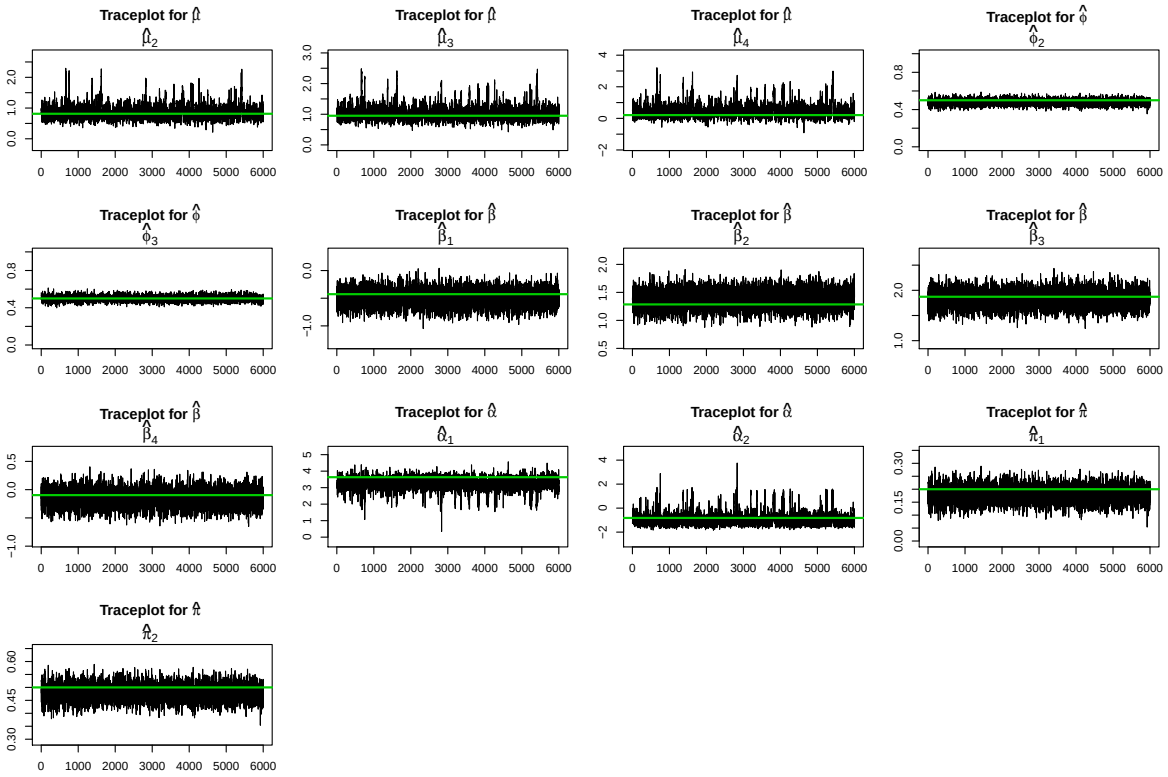

Figure H.10: ***MH simulation study***: *Trace plots of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ *when the adjacent parameters* $\phi_2 = \phi_3$. *The sample size is* $n = 500$, *the number of categories is* $q = 4$, *the number of columns is* $m = 5$ *and the number of row clusters is* $R = 4$. *The plots depict the results of the Metropolis-Hastings sampler for a sample of 6000 iterations over all the replicates* ($H = 100$) *and chains* ($S = 3$). *The green horizontal lines are the true parameter values.*

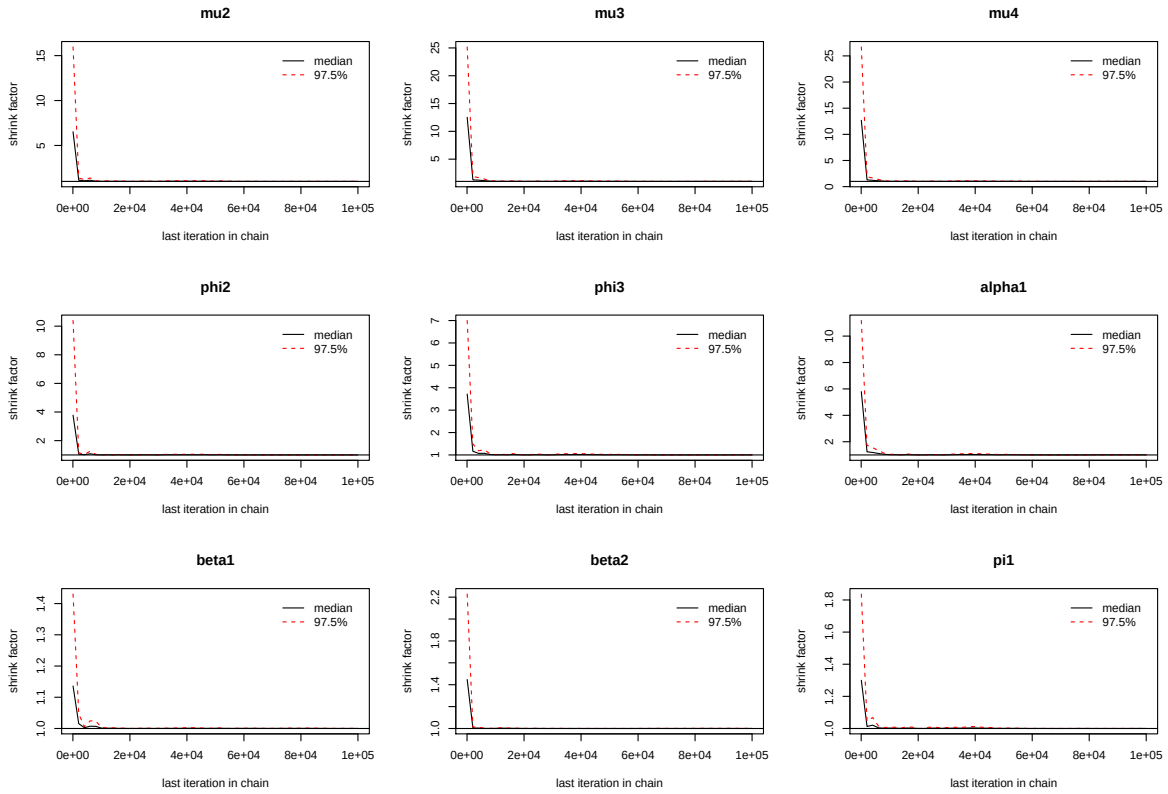

Figure H.11: ***MH simulation study***: *Evolution of Gelman and Rubin's shrink factor as the number of iterations increases. Row clustering without interaction factors* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. *The sample size is* $n = 500$, *the number of categories is* $q = 4$, *the number of columns is* $m = 3$, *and the number of row clusters is* $R = 2$. *The plots depict the results for the* $S = 3$ *chains for one replicate. For each parameter, the shrink factor converges to values lower than 1.2 diagnosing that convergence is reached.*

Appendix I

Convergence Diagnostic for RJMCMC Samplers

This chapter contains an outline of the method proposed by Castelloe and Zimmerman (2002) to assess the convergence of RJMCMC samplers is described in Section I.1 and its application for two real-life data examples is illustrated in Sections I.2 and I.3.

I.1 Description of the Method

We define Ω as the vector of all the parameters to estimate, $\boldsymbol{\omega}$ as a p-vector of all the parameters retaining the same interpretation across models, ω_i is the i^{th} component of $\boldsymbol{\omega}$, and k is a parameter in Ω (but not in $\boldsymbol{\omega}$) which is an indicator of "model" (e.g. the number of components in a mixture model). Additionally, $C > 1$ is the number of chains (of all equal length T) simulated via RJMCMC with over-dispersed starting points, M is the number of distinct models visited by all of the chains, and D_{cm} is the number of times model m occurred in chain c. Thus, $\sum_{cm} D_{cm} = CT$ and $\sum_{m} D_{cm} = T \ \ \forall c$. The procedure to diagnose the convergence is outlined as follows:

1. Calculate the following estimates of variation for parameter ω_i ($i = 1, \ldots, p$):

$$
\begin{aligned}
V(\omega_i) &= \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}(\omega_{cm}^{d}-\bar{\omega}_{..}^{\cdot})^2,\\
W_c(\omega_i) &= \frac{1}{C(T-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}(\omega_{cm}^{d}-\bar{\omega}_{c.}^{\cdot})^2,\\
W_m(\omega_i) &= \frac{1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}(\omega_{cm}^{d}-\bar{\omega}_{\cdot m}^{\cdot})^2, \text{ and}\\
W_{mc}(\omega_i) &= \frac{1}{C(T-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}(\omega_{cm}^{d}-\bar{\omega}_{cm}^{\cdot})^2,
\end{aligned}
\tag{I.1}
$$

where ω_{icm}^{d} is the value of the parameter ω_i for d^{th} occurrence of model m in chain s, and

$$
\begin{aligned}
\omega_{i..}^{\cdot} &= \frac{1}{CT}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}\omega_{icm}^{d}, \qquad \omega_{icm}^{\cdot} = \frac{1}{D_{cm}}\sum_{d=1}^{D_{cm}}\omega_{icm}^{d},\\
\omega_{ic\cdot}^{\cdot} &= \frac{1}{T}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}\omega_{icm}^{d}, \quad \text{and} \quad \omega_{i\cdot m}^{\cdot} = \frac{1}{\sum_{c=1}^{C}D_{cm}}\sum_{c=1}^{C}\sum_{d=1}^{D_{cm}}\omega_{icm}^{d}.
\end{aligned}
$$

2. Calculate the equivalent estimates of variation for the vector of parameters $\boldsymbol{\omega}$ (multivariate version):

$$
\begin{aligned}
V(\boldsymbol{\omega}) &= \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}(\boldsymbol{\omega}_{cm}^{d}-\bar{\boldsymbol{\omega}}_{..}^{\cdot})(\boldsymbol{\omega}_{cm}^{d}-\bar{\boldsymbol{\omega}}_{..}^{\cdot})' && \text{(total)},\\
W_c(\boldsymbol{\omega}) &= \frac{1}{C(T-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}(\boldsymbol{\omega}_{cm}^{d}-\bar{\boldsymbol{\omega}}_{c.}^{\cdot})(\boldsymbol{\omega}_{cm}^{d}-\bar{\boldsymbol{\omega}}_{c.}^{\cdot})' && \text{(within chain)},\\
W_m(\boldsymbol{\omega}) &= \frac{1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}(\boldsymbol{\omega}_{cm}^{d}-\bar{\boldsymbol{\omega}}_{\cdot m}^{\cdot})(\boldsymbol{\omega}_{cm}^{d}-\bar{\boldsymbol{\omega}}_{\cdot m}^{\cdot})' && \text{(within model), and}\\
W_{mc}(\boldsymbol{\omega}) &= \frac{1}{C(T-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{d=1}^{D_{cm}}(\boldsymbol{\omega}_{cm}^{d}-\bar{\boldsymbol{\omega}}_{cm}^{\cdot})(\boldsymbol{\omega}_{cm}^{d}-\bar{\boldsymbol{\omega}}_{cm}^{\cdot})' && \text{(within model,}\\
& && \text{within chain)}.
\end{aligned}
\tag{I.2}
$$

3. Select a base batch size b (Brooks and Gelman (1998) recommend $b \approx \frac{T}{20}$).

4. For batches $h = 1, \ldots, \frac{T}{b}$ do the following:

   (a) Select the parameter batch in each chain:

   $$(\boldsymbol{\omega}_1^{((h-1)b+1)}, \ldots, \boldsymbol{\omega}_1^{(hb)}), \ldots, (\boldsymbol{\omega}_C^{((h-1)b+1)}, \ldots, \boldsymbol{\omega}_C^{(hb)}).$$

   (b) Compute the following set of univariate potential scale reduction factors:

   $$\text{PSRF}_1(\omega_i) = \frac{V(\omega_i)}{W_c(\omega_i)} \text{ and } \text{PSRF}_2(\omega_i) = \frac{W_m(\omega_i)}{W_{mc}(\omega_i)} \quad \text{for } i = 1, \ldots, p. \tag{I.3}$$

   (c) Compute the following set of multivariate potential scale reduction factors:

   $$\begin{aligned} &\text{MPSRF}_1(\boldsymbol{\omega}) = \text{maximum eigenvalue of } [W_c(\boldsymbol{\omega})]^{-1} V(\boldsymbol{\omega}) \text{ and} \\ &\text{MPSRF}_2(\boldsymbol{\omega}) = \text{maximum eigenvalue of } [W_{mc}(\boldsymbol{\omega})]^{-1} W_m(\boldsymbol{\omega}). \end{aligned} \tag{I.4}$$

5. Summarise the convergence diagnostic plotting:

   (a) Plot $\text{MPSRF}_1(\boldsymbol{\omega})$ and $\text{PSRF}_1(\omega_i)$ for $i = 1, \ldots, p$ vs. h.

   (b) Plot $\text{MPSRF}_2(\boldsymbol{\omega})$ and $\text{PSRF}_2(\omega_i)$ for $i = 1, \ldots, p$ vs. h.

   (c) Plot maximum eigenvalues of $V^{(h)}(\boldsymbol{\omega})$ and $W_c^{(h)}(\boldsymbol{\omega})$ together vs. h.

   (d) Plot $V^{(h)}(\omega_i)$ and $W_c^{(h)}(\omega_i)$ together vs. h, for $i = 1, \ldots, p$.

   (e) Plot maximum eigenvalues of $W_m^{(h)}(\boldsymbol{\omega})$ and $W_m W_c^{(h)}(\boldsymbol{\omega})$ together vs. h.

   (f) Plot $W_m^{(h)}(\omega_i)$ and $W_m W_c^{(h)}(\omega_i)$ together vs. h, for $i = 1, \ldots, p$.

6. Determine h_0 such that for $h \geq h_0$ the plots in Steps 5a-5b have settled close to 1, and the plots in both Steps 5c-5d and Steps 5e-5f have settled approximately to a common value.

I.2 Example 1: Applied Statistics Course Feedback Forms

The convergence diagnostic plots for the Applied Statistics course feedback data set are shown in Figures I.1-I.4. Figures I.1 and I.2 imply that convergence is likely to have occurred by the 7^{th} batch where the MPSRF's and PSRF's stay below 1.02 for all the fixed-dimensional parameters. Figures I.3 and I.4 show that pairs of plots for both Steps 5c-5d and Steps 5e-5f stay very close together throughout.

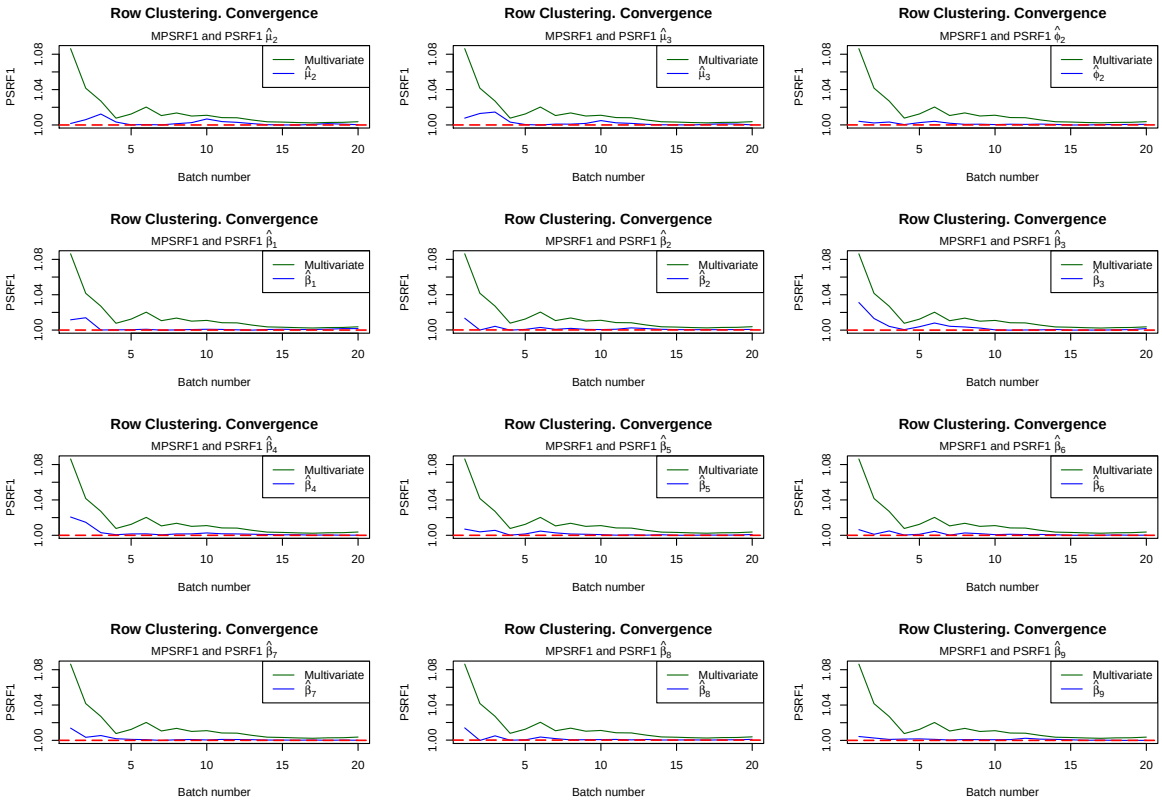

Figure I.1: ***Applied Statistics course feedback forms data set:*** *Potential scale reduction factor (see $PSRF_1$ in eq. (I.3) and $MPSRF_1$ in eq. (I.4)) plots for fixed-dimensional parameters $\{\mu_k\}$, $\{\phi_k\}$ and $\{\beta_j\}$ for the row clustering model. Five RJMCMC chains were used and $h = 20$ batches were generated. The blue line in each plot corresponds to the $PSRF_1$ for one particular parameter. The green line is the multivariate $MPSRF_1$ version and is the same throughout all the plots. The convergence is likely to have occurred by the 7^{th} batch where the $MPSRF_1$ and $PSRF_1$ stay below 1.02 for all the fixed-dimensional parameters.*

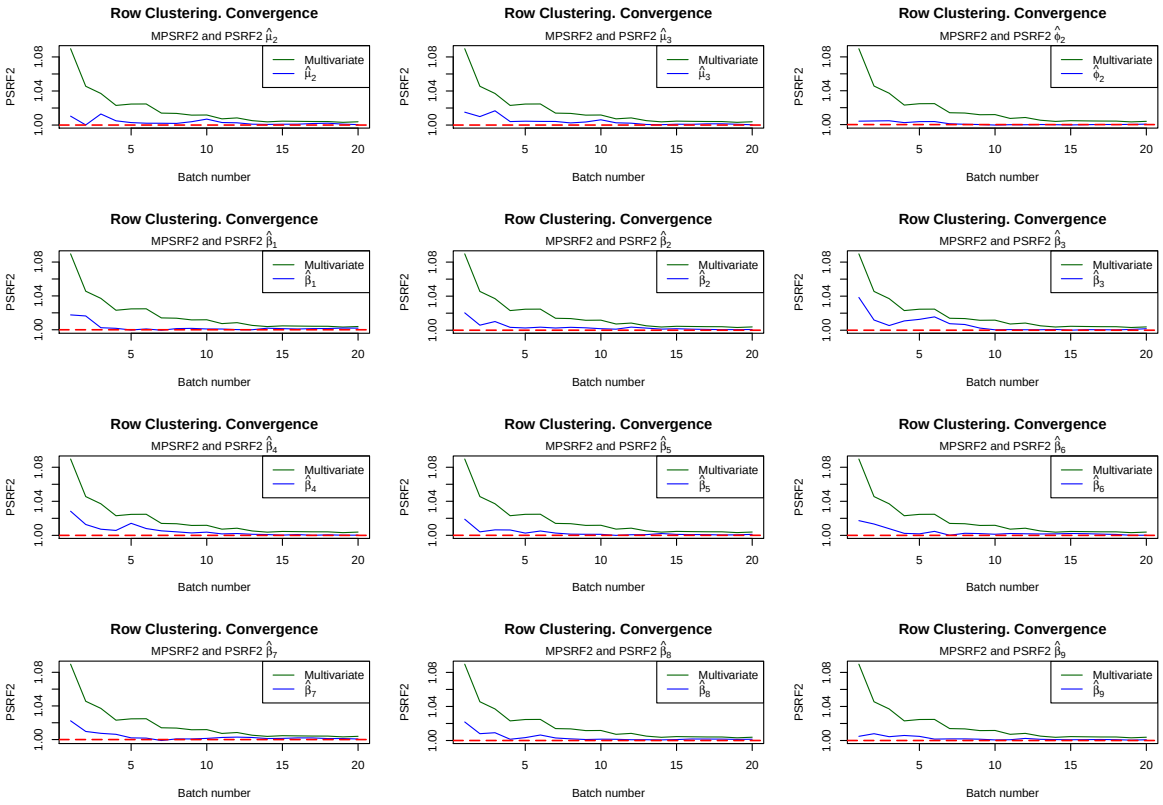

Figure I.2: ***Applied Statistics course feedback forms data set:*** *Potential scale reduction factor (see $PSRF_2$ in eq. (I.3) and $MPSRF_2$ in eq. (I.4)) plots for fixed-dimensional parameters $\{\mu_k\}$, $\{\phi_k\}$ and $\{\beta_j\}$ for the row clustering model. Five RJMCMC chains were used and $h = 20$ batches were generated. The blue line in each plot corresponds to the $PSRF_2$ for one particular parameter. The green line is the multivariate $MPSRF_2$ version and is the same throughout all the plots. The convergence is likely to have occurred by the 7^{th} batch where the $MPSRF_2$ and $PSRF_2$ stay below 1.02 for all the fixed-dimensional parameters.*

I.3 Example 2: Spider Data

The convergence diagnostic plots for the Spider data set are shown in Figures I.5-I.8. Figures I.5 and I.6 imply that convergence is likely to have occurred by the 9^{th} batch where the MPSRF's and PSRF's stay below 1.02 for all the fixed-dimensional parameters. Figures I.7 and I.8 show that pairs of plots for both Steps 5c-5d and Steps 5e-5f stay very close together throughout.

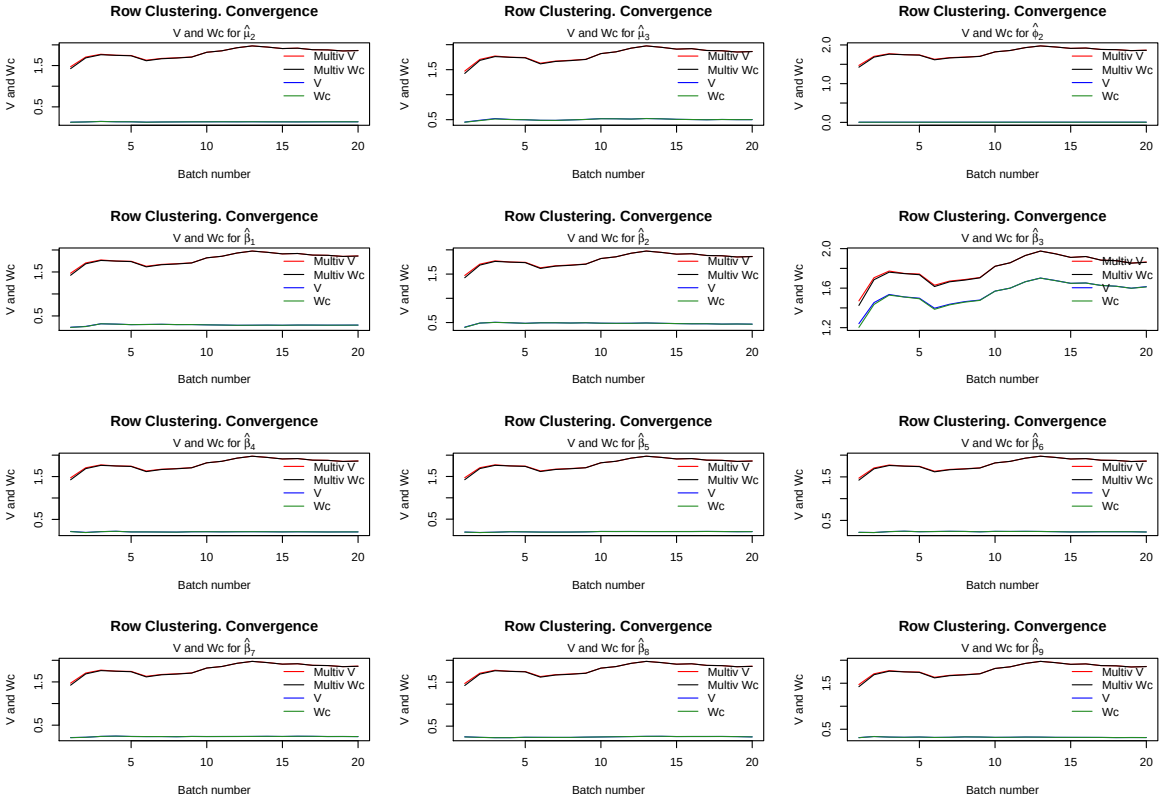

Figure I.3: ***Applied Statistics course feedback forms data set:*** *Plots of $V^{(h)}(\omega_i)$, $W_c^{(h)}(\omega_i)$ (see eq. (I.1)) and the maximum eigenvalues of $V^{(h)}(\boldsymbol{\omega})$ and $W_c^{(h)}(\boldsymbol{\omega})$ (see eq. (I.2)) for fixed-dimensional parameters $\{\mu_k\}$, $\{\phi_k\}$ and $\{\beta_j\}$ for row clustering model. Five RJMCMC chains were used and $h = 20$ batches were generated. The red line corresponds to $V^{(h)}(\boldsymbol{\omega})$, the black line to $W_c^{(h)}(\boldsymbol{\omega})$, the blue line to $V^{(h)}(\omega_i)$, and the green line to $W_c^{(h)}(\omega_i)$. The plots imply convergence as the pair $V^{(h)}(\omega_i)$ and $W_c^{(h)}(\omega_i)$ and also the pair of the maximum eigenvalues of $V^{(h)}(\boldsymbol{\omega})$ and $W_c^{(h)}(\boldsymbol{\omega})$ stay very close together throughout the batches.*

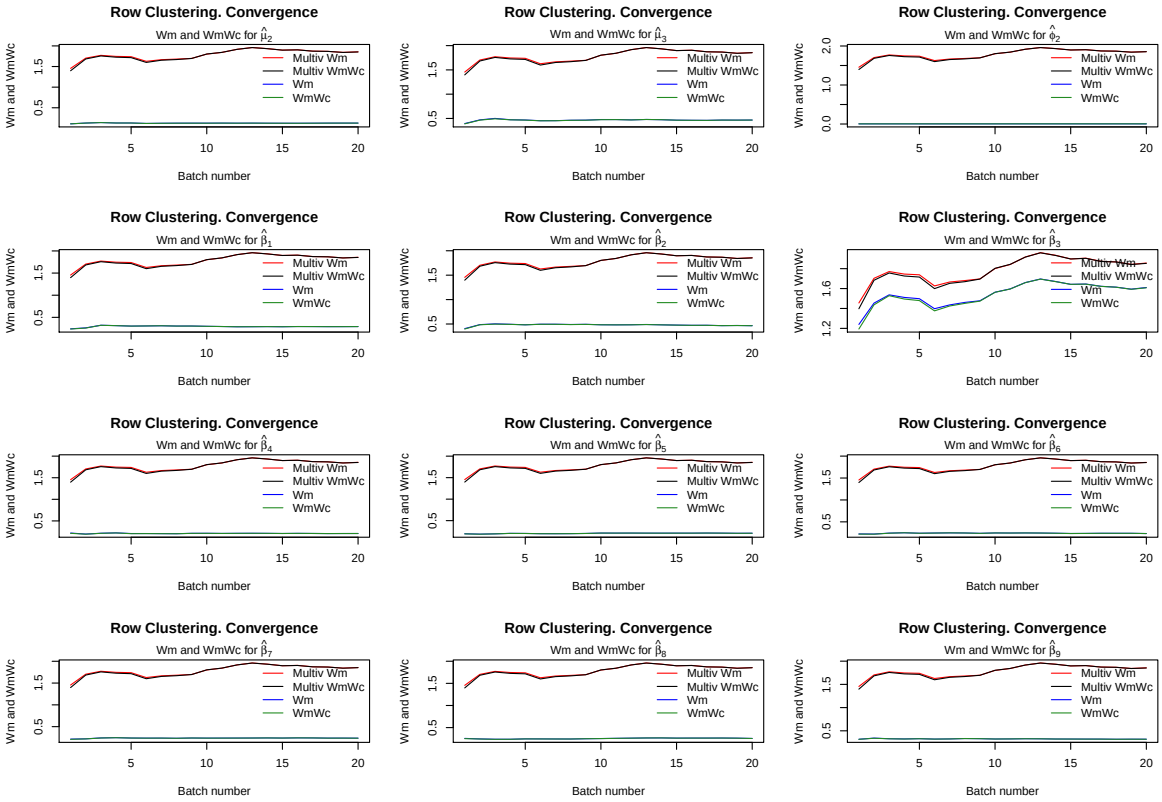

Figure I.4: ***Applied Statistics course feedback forms data set:*** *Plots of $W_m^{(h)}(\omega_i)$, $W_m W_c^{(h)}(\omega_i)$ (see eq. (I.1)) and the maximum eigenvalues of $W_m^{(h)}(\boldsymbol{\omega})$ and $W_m W_c^{(h)}(\boldsymbol{\omega})$ (see eq. (I.2)) for fixed-dimensional parameters $\{\mu_k\}$, $\{\phi_k\}$ and $\{\beta_j\}$ for row clustering model. Five RJMCMC chains were used and $h = 20$ batches were generated. The red line corresponds to $W_m^{(h)}(\boldsymbol{\omega})$, the black line to $W_m W_c^{(h)}(\boldsymbol{\omega})$, the blue line to $W_m^{(h)}(\omega_i)$, and the green line to $W_m W_c^{(h)}(\omega_i)$. The plots imply convergence as the pair $V^{(h)}(\omega_i)$ and $W_c^{(h)}(\omega_i)$ and also the pair of the maximum eigenvalues of $V^{(h)}(\boldsymbol{\omega})$ and $W_c^{(h)}(\boldsymbol{\omega})$ stay very close together after the 7^{th} batch.*

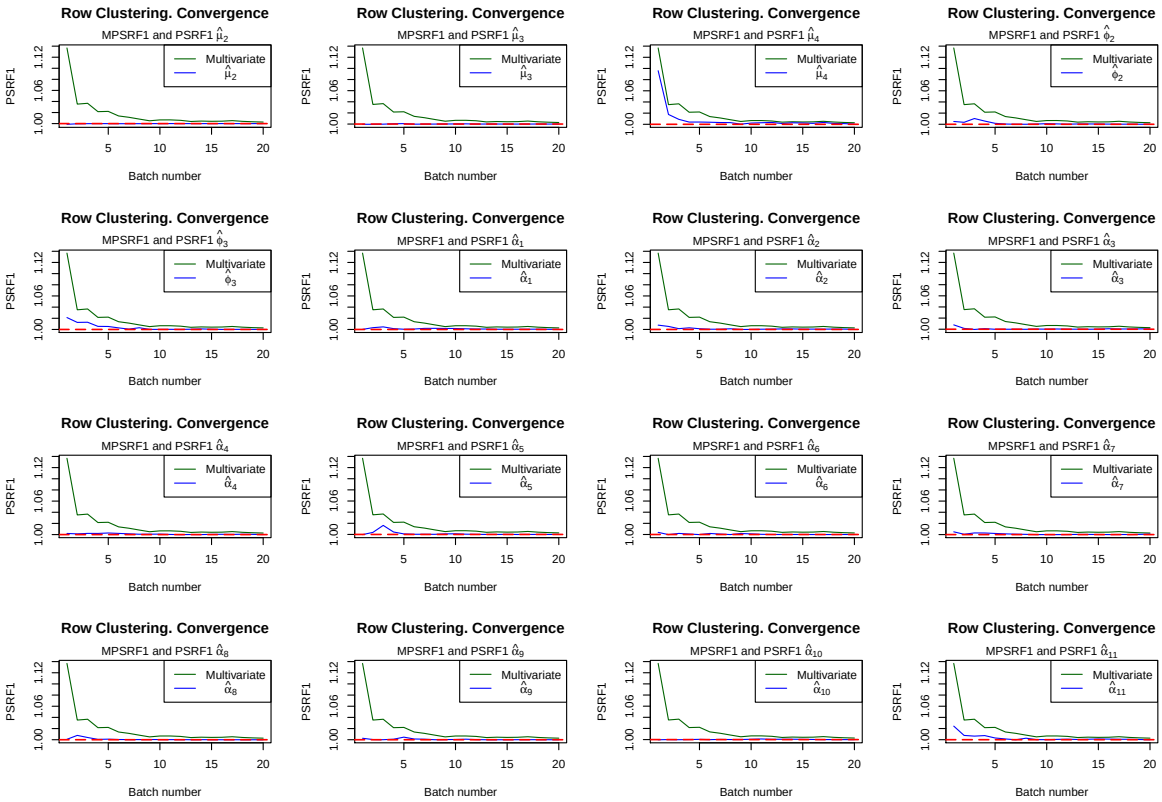

Figure I.5: ***Spider data set:*** *Potential scale reduction factor (see $PSRF_1$ in eq. (I.3) and $MPSRF_1$ in eq. (I.4)) plots for fixed-dimensional parameters $\{\mu_k\}$, $\{\phi_k\}$ and $\{\beta_j\}$ for row clustering model. Five RJMCMC chains were used and $h = 20$ batches were generated. The blue line in each plot corresponds to the $PSRF_1$ for one particular parameter. The green line is the multivariate $MPSRF_1$ version and is the same throughout all the plots. The convergence is likely to have occurred by the 9^{th} batch where the $MPSRF_1$ and $PSRF_1$ stay below 1.02 for all the fixed-dimensional parameters.*

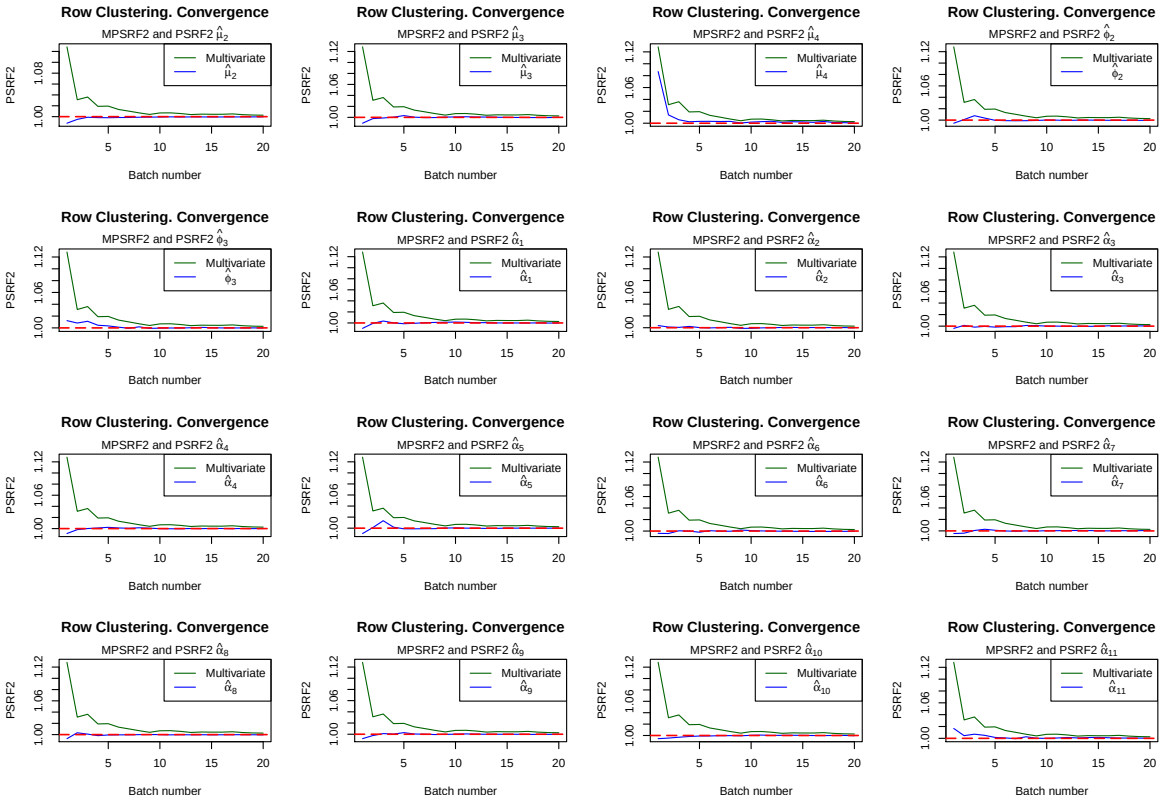

Figure I.6: ***Spider data set:*** *Potential scale reduction factor (see $PSRF_2$ in eq. (I.3) and $MPSRF_2$ in eq. (I.4)) plots for fixed-dimensional parameters $\{\mu_k\}$, $\{\phi_k\}$ and $\{\beta_j\}$ for row clustering model. Five RJMCMC chains were used and $h = 20$ batches were generated. The blue line in each plot corresponds to the $PSRF_2$ for one particular parameter. The green line is the multivariate $MPSRF_2$ version and is the same throughout all the plots. The convergence is likely to have occurred by the 9^{th} batch where the $MPSRF_2$ and $PSRF_2$ stay below 1.02 for all the fixed-dimensional parameters.*

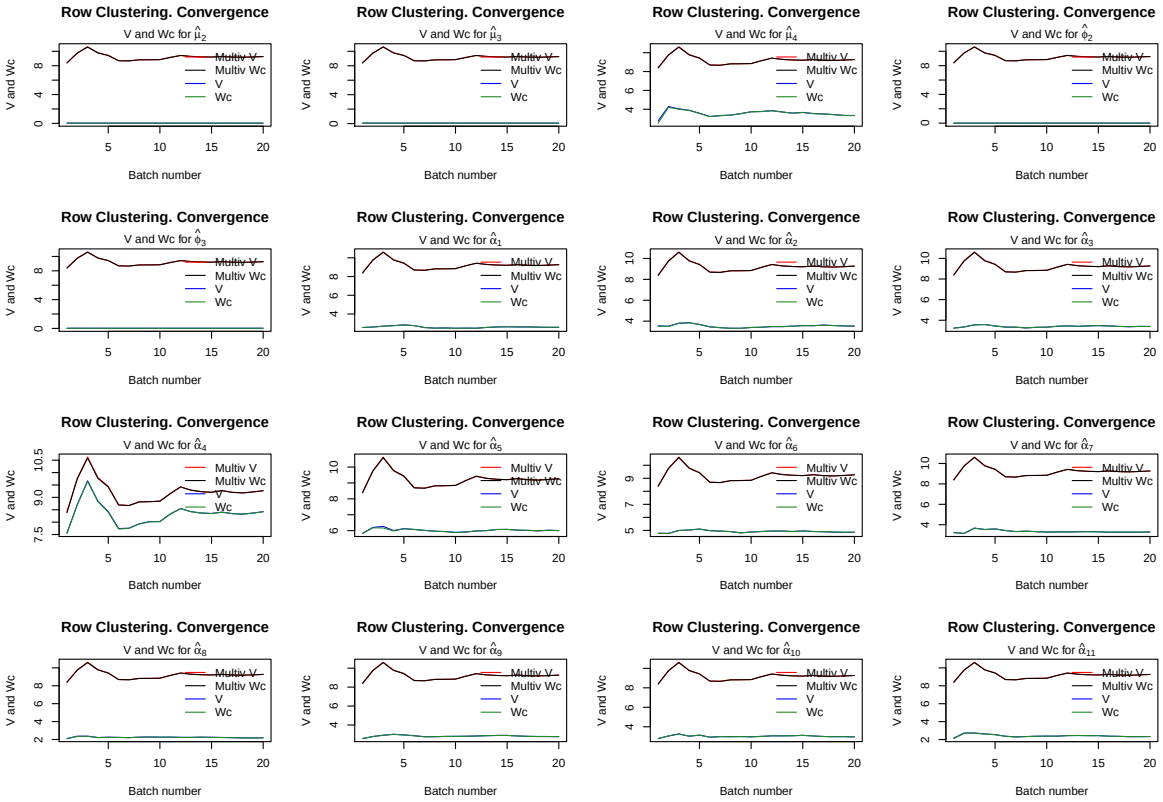

Figure I.7: ***Spider data set:*** *Plots of $V^{(h)}(\omega_i)$, $W_c^{(h)}(\omega_i)$ (see eq. (I.1)) and the maximum eigenvalues of $V^{(h)}(\boldsymbol{\omega})$ and $W_c^{(h)}(\boldsymbol{\omega})$ (see eq. (I.2)) for fixed-dimensional parameters $\{\mu_k\}$, $\{\phi_k\}$ and $\{\alpha_i\}$ for column clustering model. Five RJMCMC chains were used and $h = 20$ batches were generated. The red line corresponds to $V^{(h)}(\boldsymbol{\omega})$, the black line to $W_c^{(h)}(\boldsymbol{\omega})$, the blue line to $V^{(h)}(\omega_i)$, and the green line to $W_c^{(h)}(\omega_i)$. The plots imply convergence as the pair $V^{(h)}(\omega_i)$ and $W_c^{(h)}(\omega_i)$ and also the pair of the maximum eigenvalues of $V^{(h)}(\boldsymbol{\omega})$ and $W_c^{(h)}(\boldsymbol{\omega})$ stay very close together throughout the batches.*

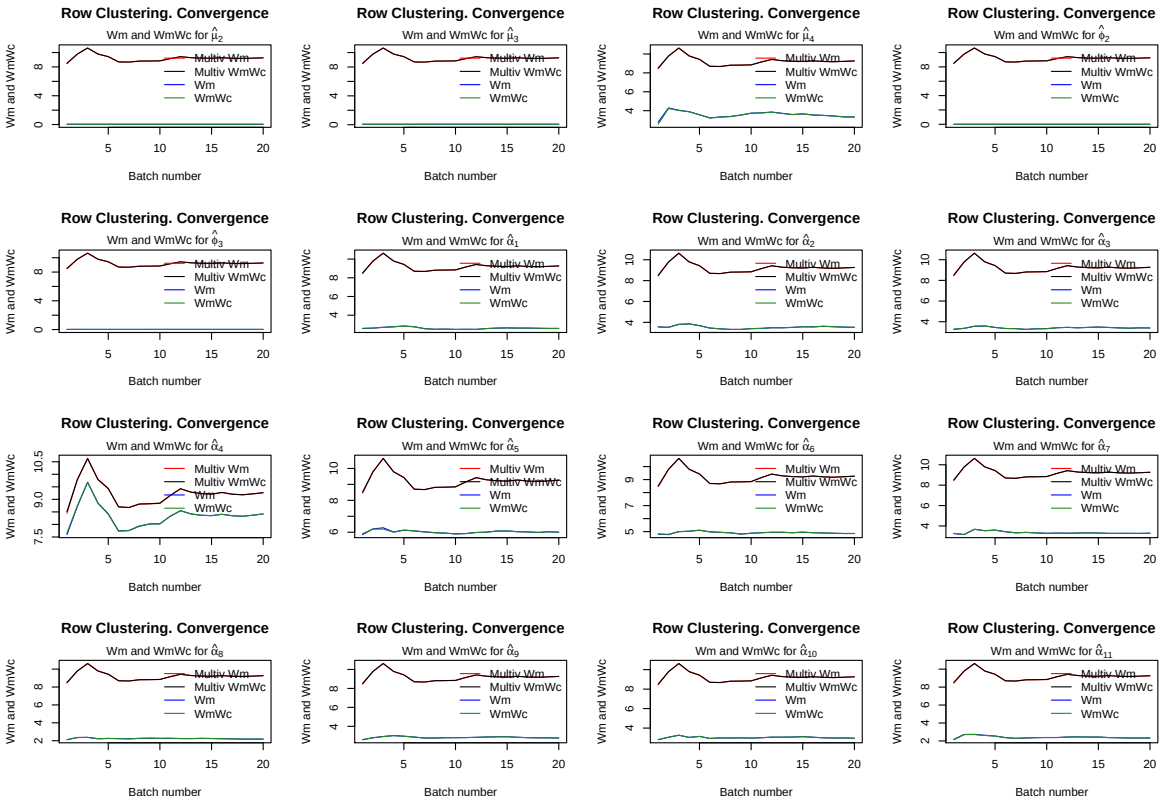

Figure I.8: ***Spider data set:*** *Plots of $W_m^{(h)}(\omega_i)$, $W_m W_c^{(h)}(\omega_i)$ (see eq. (I.1)) and the maximum eigenvalues of $W_m^{(h)}(\boldsymbol{\omega})$ and $W_m W_c^{(h)}(\boldsymbol{\omega})$ (see eq. (I.2)) for fixed-dimensional parameters $\{\mu_k\}$, $\{\phi_k\}$ and $\{\alpha_i\}$ for column clustering model. Five RJMCMC chains were used and $h = 20$ batches were generated. The red line corresponds to $W_m^{(h)}(\boldsymbol{\omega})$, the black line to $W_m W_c^{(h)}(\boldsymbol{\omega})$, the blue line to $W_m^{(h)}(\omega_i)$, and the green line to $W_m W_c^{(h)}(\omega_i)$. The plots imply convergence as the pair $V^{(h)}(\omega_i)$ and $W_c^{(h)}(\omega_i)$ and also the pair of the maximum eigenvalues of $V^{(h)}(\boldsymbol{\omega})$ and $W_c^{(h)}(\boldsymbol{\omega})$ stay very close together throughout the batches.*

Appendix J

RJMCMC Sampler. Simulation Study

In this appendix, the results of the simulation study of the RJMCMC sampler for one-dimensional clustering are shown. Given the number of replicates (data sets) H and the number of chains S, Figures J.1-J.4 show the $HS = 100 \times 10 = 1000$ separate MAP estimators of all the parameters taken in pairs and plotted against each other for the row clustering model with $R = 3, \ldots, 6$ row clusters respectively. The red diamond point represents the true value of the parameter. The MAP estimators are around the true value of the parameter in all the scatter plots. Figures J.5-J.12 show the marginal posterior distribution and trace plot for all the parameters for $R = 3, \ldots, 6$ number of row clusters respectively. The expected values of the posterior distribution are very close to the true values (green vertical lines) and the 95% HPD credible interval includes the true parameter values in all the cases. The trace plots on Figures J.6-J.12 show a good mixing on all the parameters. Finally, Table J.1 shows the proportion of times across the HS possible chains where the 95% HPD region includes the true value of the parameters with variable dimension ($\{\alpha_r\}$ and $\{\pi_r\}$). The proportion of times that true parameters are covered by the 95% HPD is 90% at least which is very satisfactory.

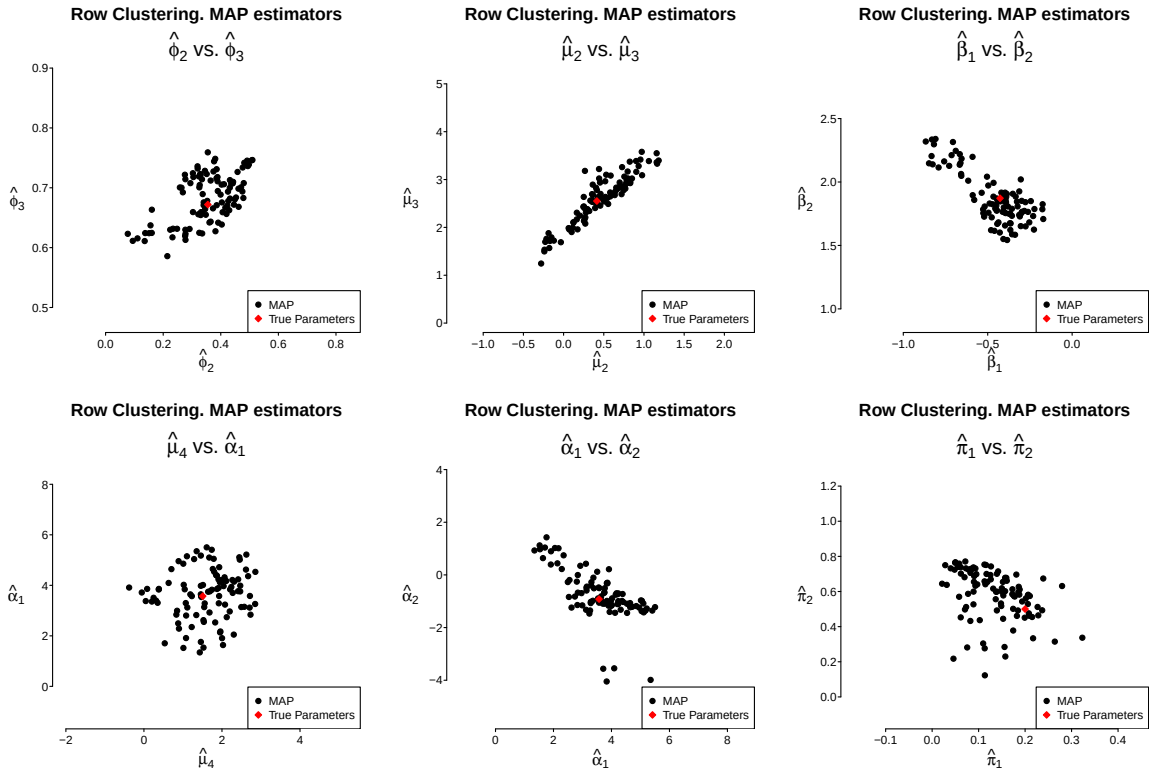

Figure J.1: ***RJMCMC simulation study R=3:*** *Scatter plots depicting the maximum a posteriori estimator (MAP) across all the replicates ($H = 100$) and chains ($S = 10$) for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ with $R = 3$ row clusters. The sample size for each chain and replica is $n = 1000$, the number of categories is $q = 4$, and the number of columns is $m = 3$. The black points are the MAP estimators and the red diamond point represents the true value of the parameter. The plots comparing adjacent pair of $\{\widehat{\mu}_k\}$ parameters show its correlation. The results are centered on the true parameters.*

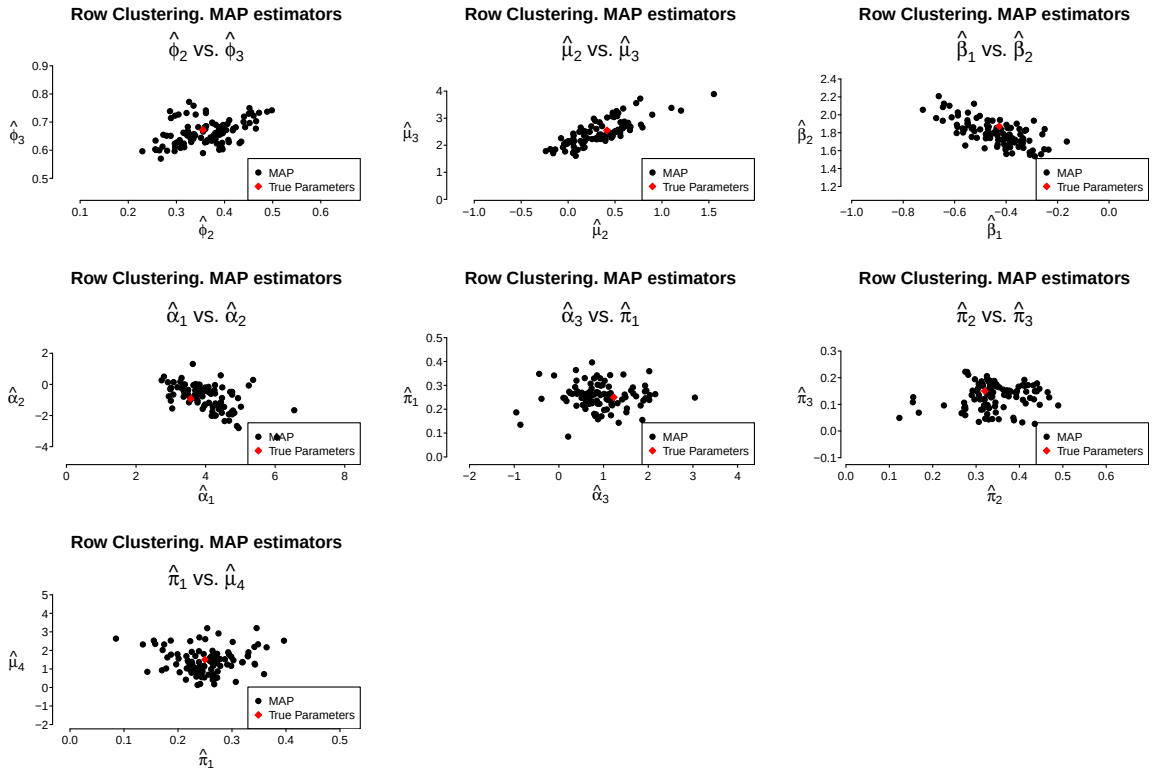

Figure J.2: ***RJMCMC simulation study R=4:*** *Scatter plots depicting the maximum a posteriori estimator (MAP) across all the replicates ($H = 100$) and chains ($S = 10$) for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ with $R = 4$ row clusters. The sample size for each chain and replica is $n = 1000$, the number of categories is $q = 4$, and the number of columns is $m = 3$. The black points are the MAP estimators and the red diamond point represents the true value of the parameter. The plots comparing adjacent pair of $\{\widehat{\mu}_k\}$ parameters show its correlation. The results are centered on the true parameters.*

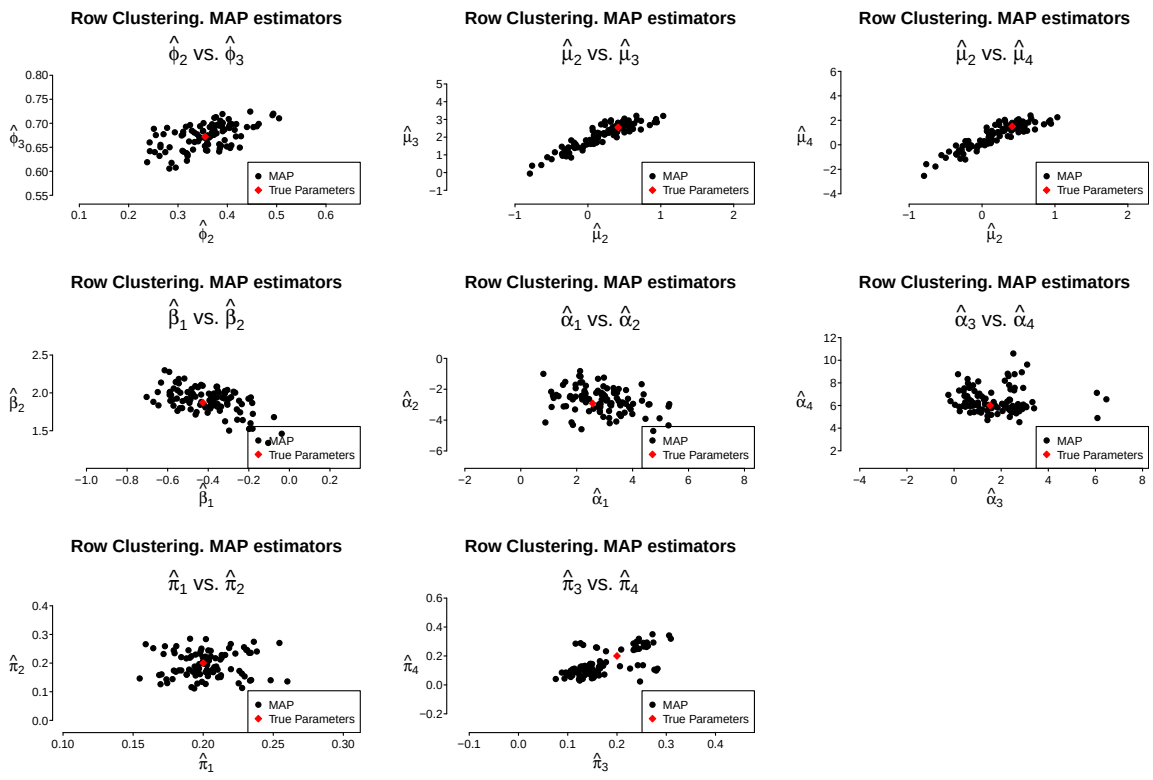

Figure J.3: ***RJMCMC simulation study R=5:*** *Scatter plots depicting the maximum a posteriori estimator (MAP) across all the replicates ($H = 100$) and chains ($S = 10$) for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ with $R = 5$ row clusters. The sample size for each chain and replica is $n = 1000$, the number of categories is $q = 4$, and the number of columns is $m = 3$. The black points are the MAP estimators and the red diamond point represents the true value of the parameter. The plots comparing adjacent pair of $\{\widehat{\mu}_k\}$ parameters show its correlation. The plots comparing a pair of $\{\widehat{\pi}_r\}$ parameters show label switching (inverse correlation of $\{\widehat{\pi}_r\}$ values). The results are centered on the true parameters.*

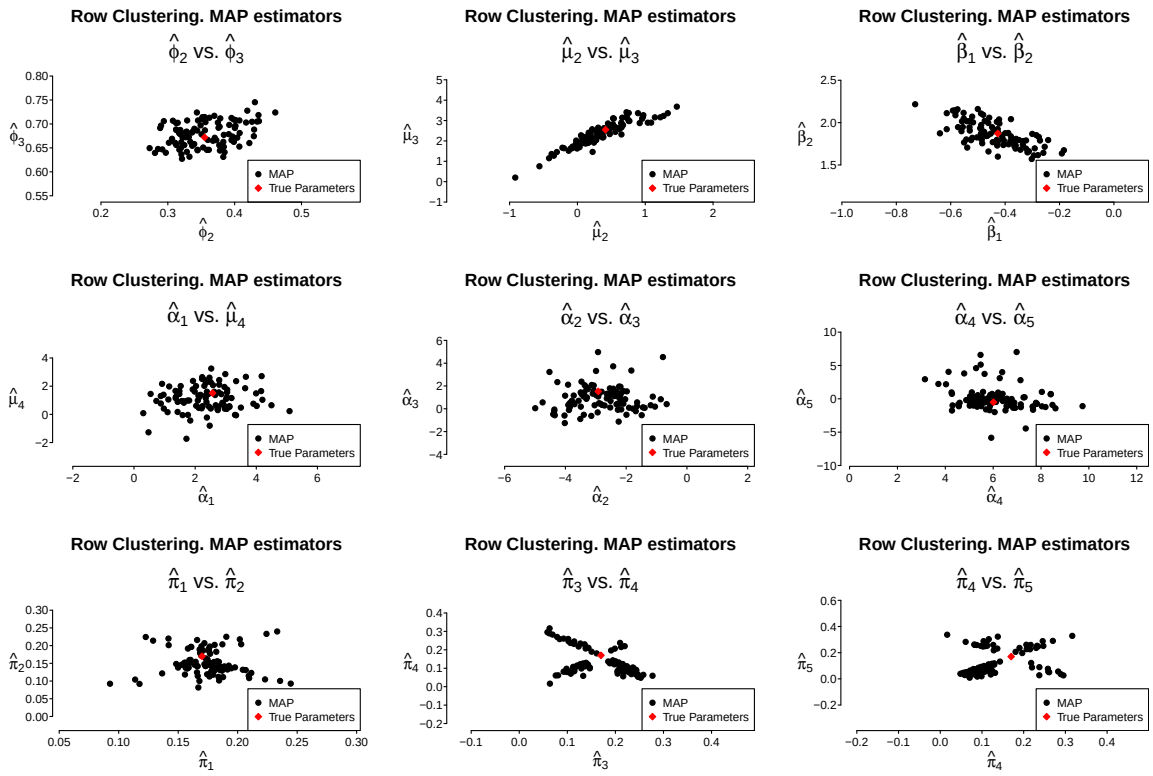

Figure J.4: ***RJMCMC simulation study R=6:*** *Scatter plots depicting the maximum a posteriori estimator (MAP) across all the replicates ($H = 100$) and chains ($S = 10$) for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$ with $R = 6$ row clusters. The sample size for each chain and replica is $n = 1000$, the number of categories is $q = 4$, and the number of columns is $m = 3$. The black points are the MAP estimators and the red diamond point represents the true value of the parameter. The plots comparing adjacent pair of $\{\widehat{\mu}_k\}$ parameters show its correlation. The plots comparing a pair of $\{\widehat{\pi}_r\}$ parameters show label switching (inverse correlation of $\{\widehat{\pi}_r\}$ values). The results are centered on the true parameters.*

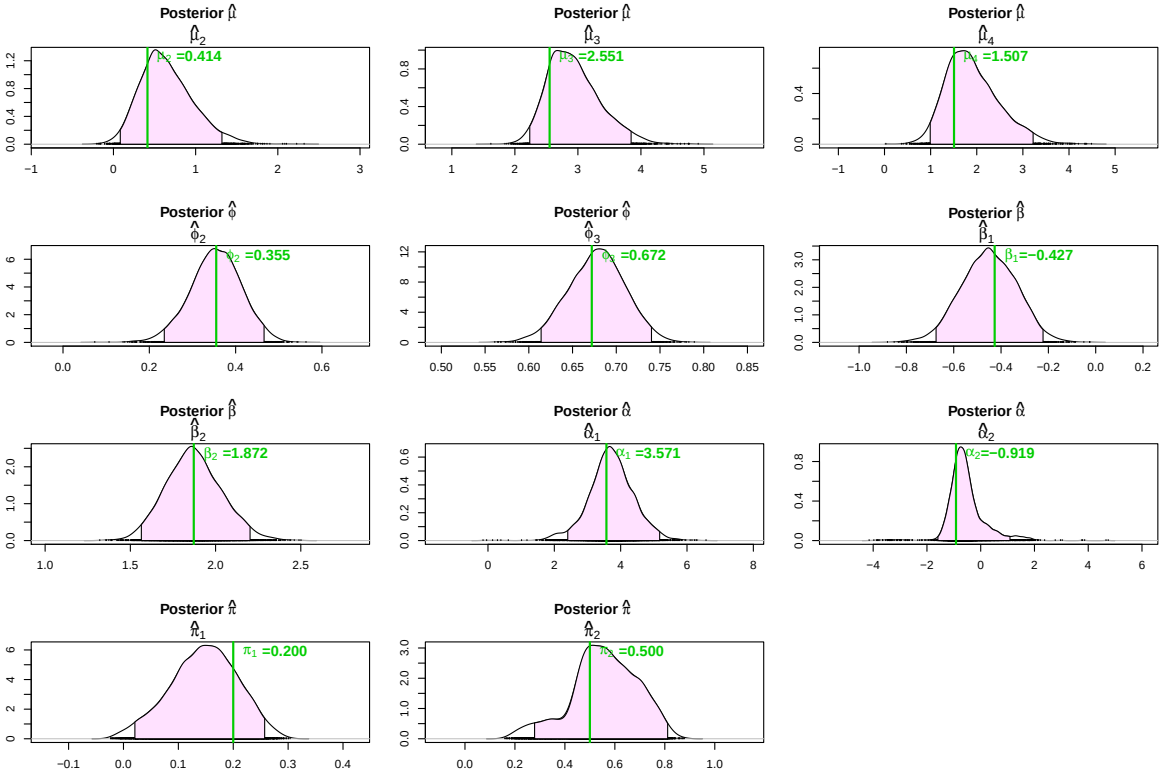

Figure J.5: ***RJMCMC simulation study R=3:*** *Density plot of the parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The sample size is $n = 1000$, the number of categories is $q = 4$, the number of columns is $m = 3$ and the number of row clusters are $R = 3$. The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates ($H = 100$) and chains ($S = 10$). The green vertical lines are the true parameter value and 95% HPD credible intervals are shown with shading area.*

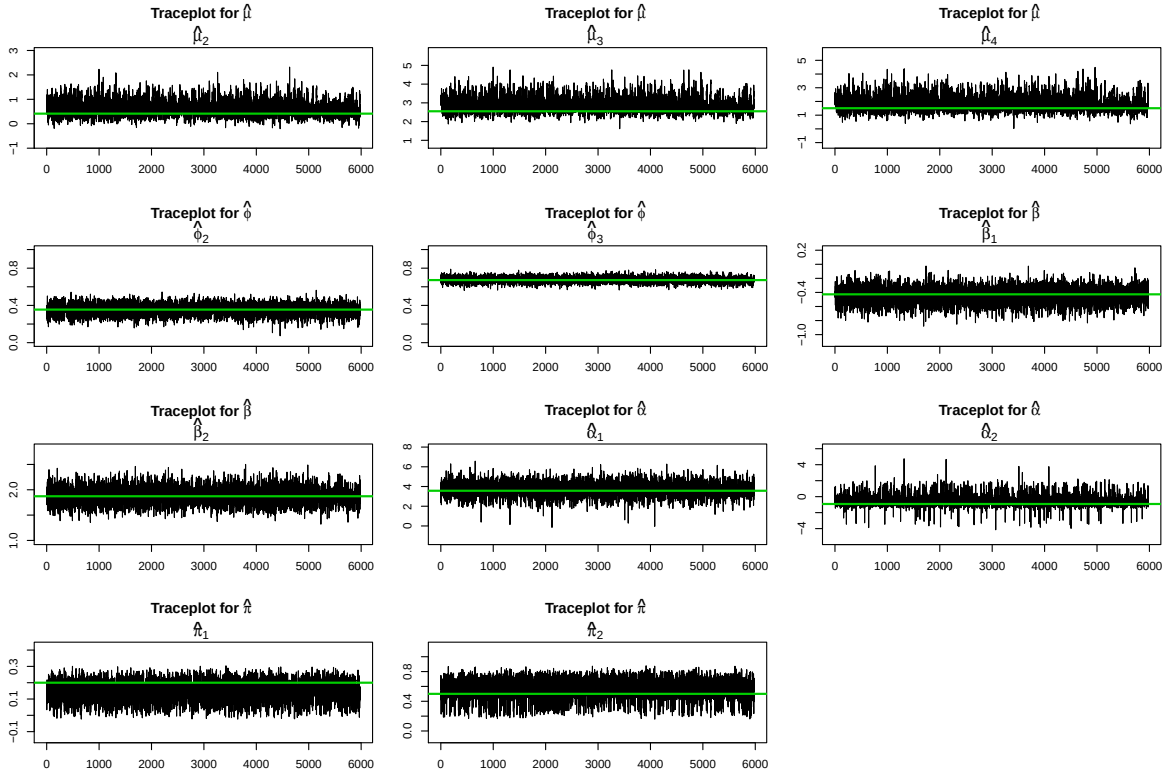

Figure J.6: ***RJMCMC simulation study R=3:*** *Trace plot of the parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The sample size is $n = 1000$, the number of categories is $q = 4$, the number of columns is $m = 3$ and the number of row clusters are $R = 3$. The plots depict the results of the RJMCMC sampler for a sample of 6000 iterations over all the replicates ($H = 100$) and chains ($S = 10$). The green horizontal lines are the true parameter value.*

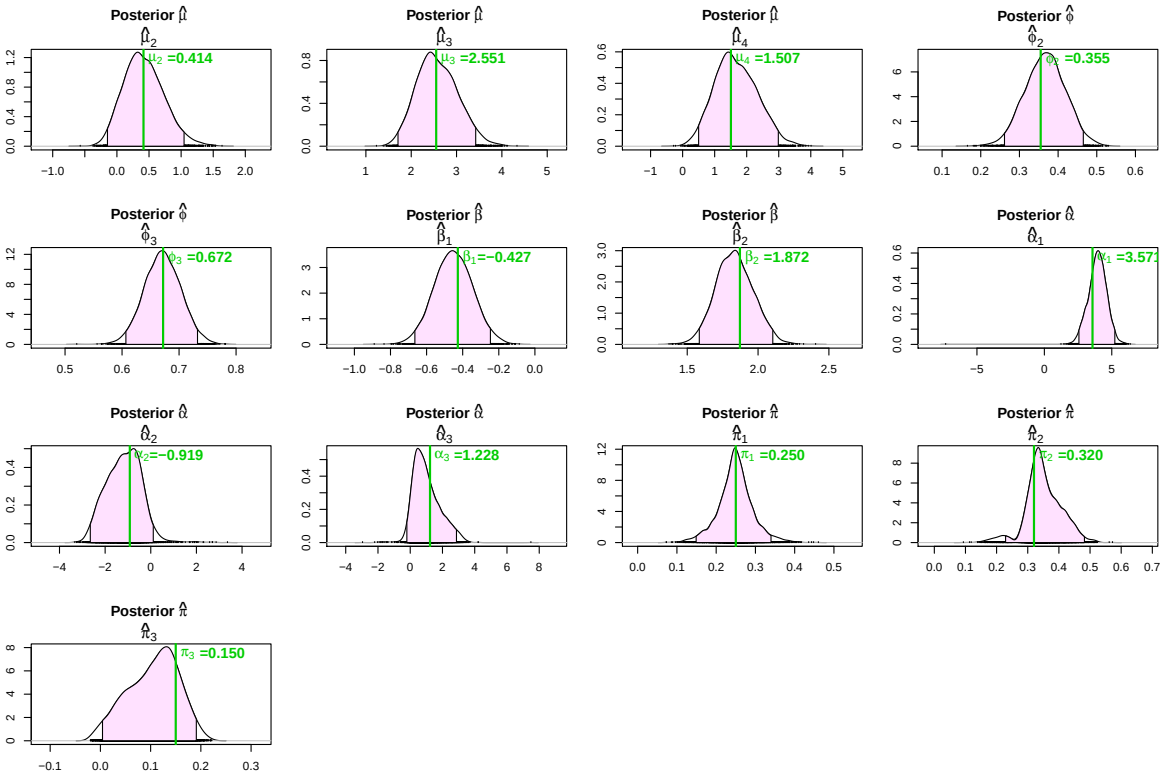

Figure J.7: ***RJMCMC simulation study R=4:*** *Density plot of the parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The sample size is $n = 1000$, the number of categories is $q = 4$, the number of columns is $m = 3$ and the number of row clusters are $R = 4$. The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates ($H = 100$) and chains ($S = 10$). The green vertical lines are the true parameter value and 95% HPD credible intervals are shown with shading area.*

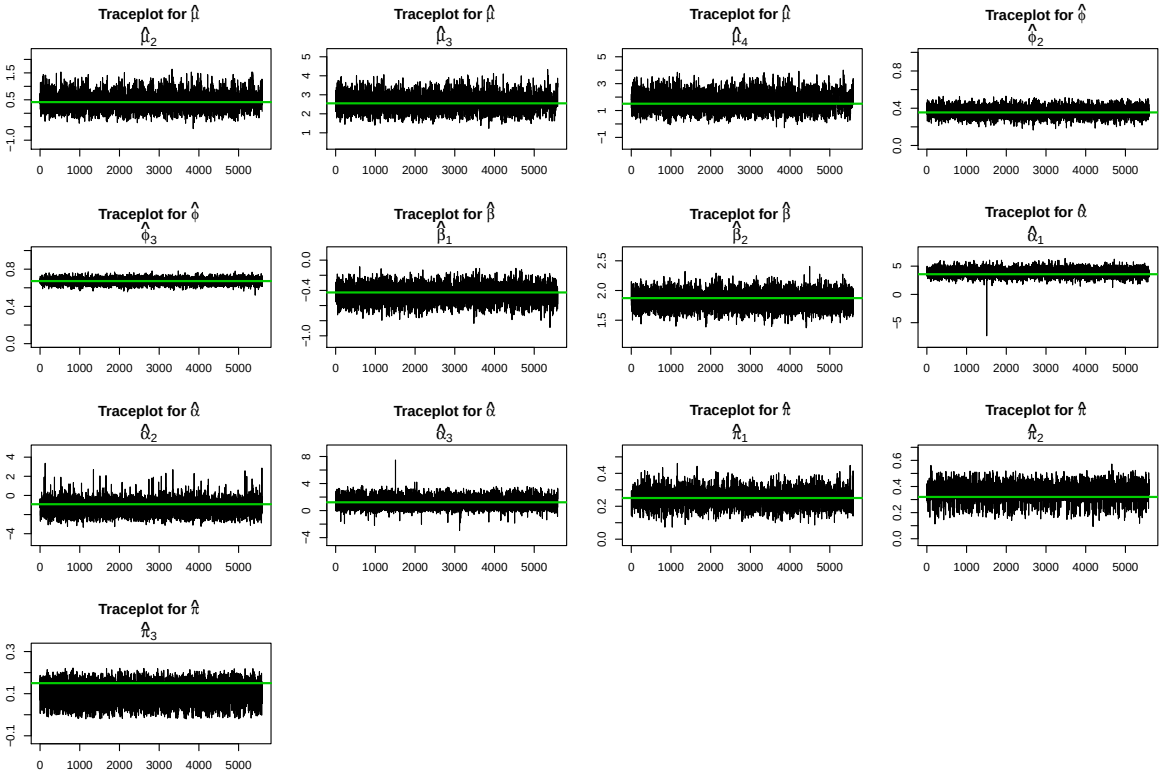

Figure J.8: ***RJMCMC simulation study R=4:*** *Trace plot of the parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The sample size is $n = 1000$, the number of categories is $q = 4$, the number of columns is $m = 3$ and the number of row clusters are $R = 4$. The plots depict the results of the RJMCMC sampler for a sample of 6000 iterations over all the replicates ($H = 100$) and chains ($S = 10$). The green horizontal lines are the true parameter value.*

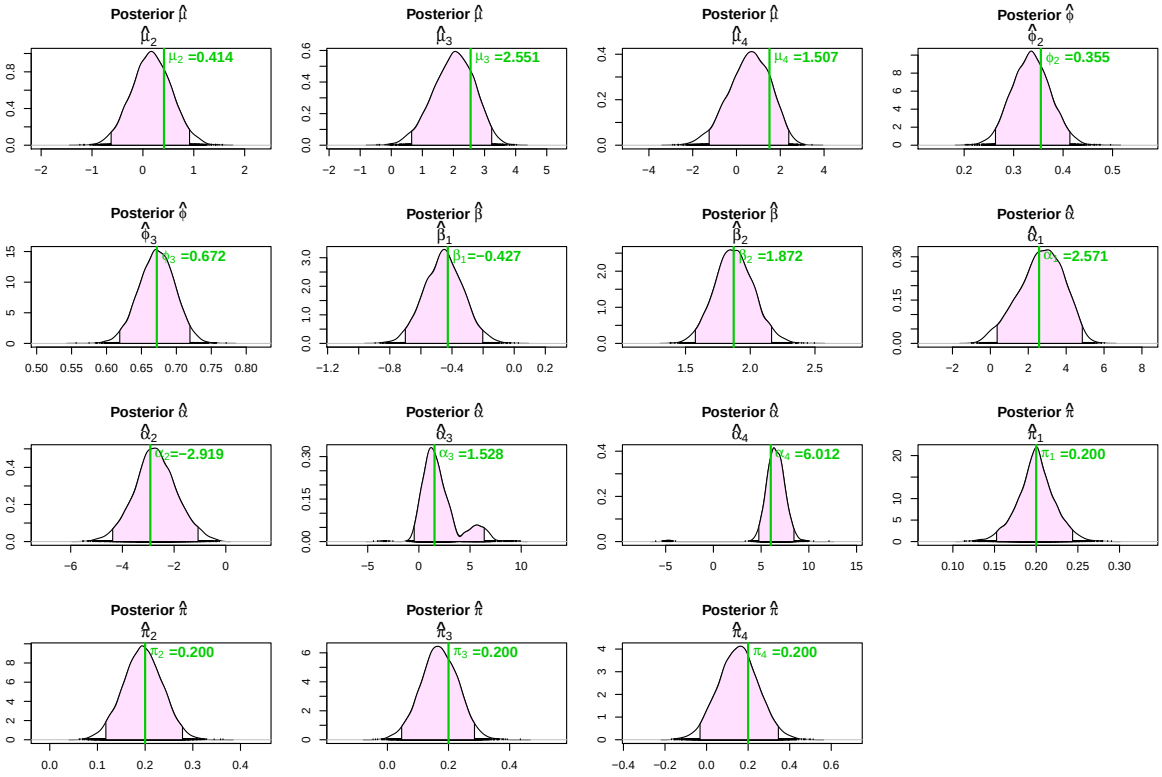

Figure J.9: ***RJMCMC simulation study R=5:*** *Density plot of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$*. The sample size is* $n = 1000$*, the number of categories is* $q = 4$*, the number of columns is* $m = 3$ *and the number of row clusters are* $R = 5$*. The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 10$*). The green vertical lines are the true parameter value and 95% HPD credible intervals are shown with shading area.*

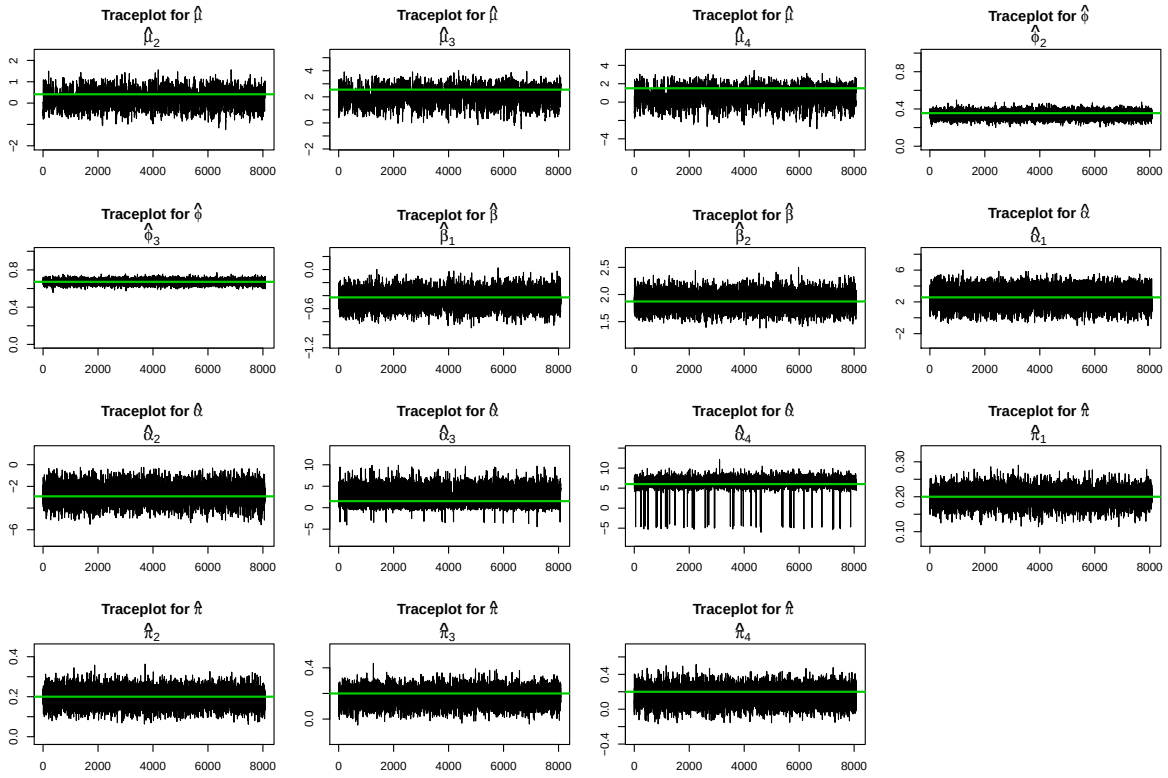

Figure J.10: ***RJMCMC simulation study R=5:*** *Trace plot of the parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The sample size is $n = 1000$, the number of categories is $q = 4$, the number of columns is $m = 3$ and the number of row clusters are $R = 5$. The plots depict the results of the RJMCMC sampler for a sample of 6000 iterations over all the replicates ($H = 100$) and chains ($S = 10$). The green horizontal lines are the true parameter value. Jumps in the trace plot of $\widehat{\alpha}_4$ indicates that label switching problem is occurring.*

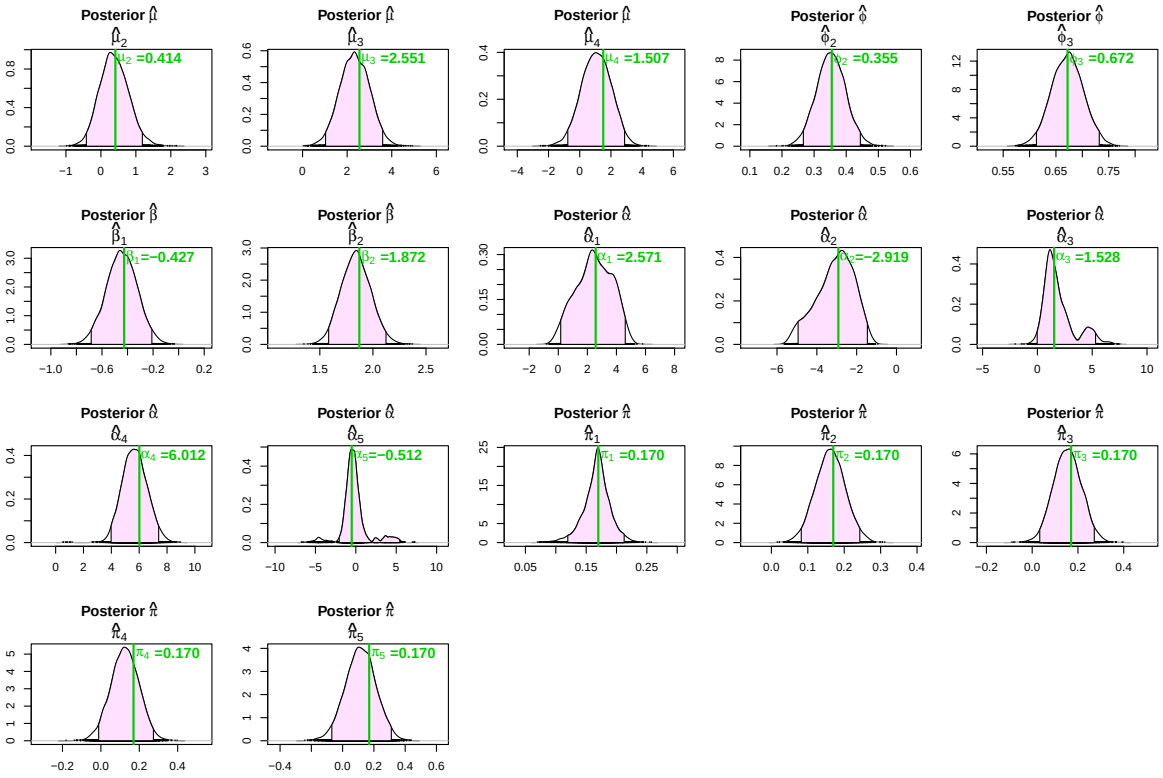

Figure J.11: ***RJMCMC simulation study R=6:*** *Density plot of the parameters for stereotype model including row clustering* $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$*. The sample size is* $n = 1000$*, the number of categories is* $q = 4$*, the number of columns is* $m = 3$ *and the number of row clusters are* $R = 6$*. The density plots depict the marginal posterior distribution for a sample of 6000 iterations over all the replicates (*$H = 100$*) and chains (*$S = 10$*). The green vertical lines are the true parameter value and 95% HPD credible intervals are shown with shading area.*

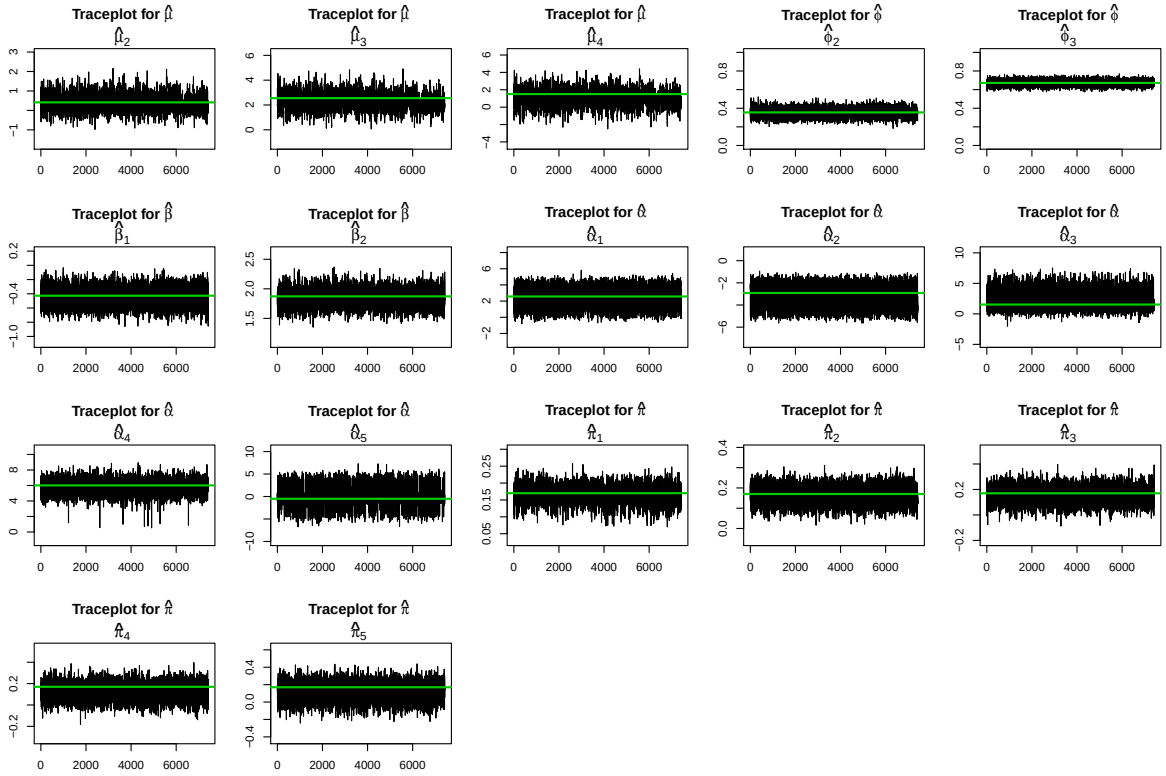

Figure J.12: ***RJMCMC simulation study R=6:*** *Trace plot of the parameters for stereotype model including row clustering $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$. The sample size is $n = 1000$, the number of categories is $q = 4$, the number of columns is $m = 3$ and the number of row clusters are $R = 6$. The plots depict the results of the RJMCMC sampler for a sample of 6000 iterations over all the replicates ($H = 100$) and chains ($S = 10$). The green horizontal lines are the true parameter value.*

Table J.1: **RJMCMC simulation study:** Proportion of times the 95% HPD region includes the true value of the variable-dimensional parameters across the HS possible chains for the for the stereotype model including row clustering without interaction factors $\boldsymbol{\mu_k} + \boldsymbol{\phi_k}(\boldsymbol{\alpha_r} + \boldsymbol{\beta_j})$.

R	True parameters	Proportion of times within 95% HPD region
2	$\alpha_1 = 3.571$	91%
	$\pi_1 = 0.350$	94%
3	$\alpha_1 = 3.571$	94%
	$\alpha_2 = -0.919$	96%
	$\pi_1 = 0.200$	95%
	$\pi_1 = 0.500$	91%
4	$\alpha_1 = 3.571$	92%
	$\alpha_2 = -0.919$	91%
	$\alpha_3 = 1.228$	95%
	$\pi_1 = 0.250$	92%
	$\pi_2 = 0.320$	93%
	$\pi_3 = 0.150$	94%
5	$\alpha_1 = 2.571$	95%
	$\alpha_2 = -2.919$	90%
	$\alpha_3 = 1.528$	91%
	$\alpha_4 = 6.012$	96%
	$\pi_1 = 0.200$	91%
	$\pi_2 = 0.200$	97%
	$\pi_3 = 0.200$	91%
	$\pi_4 = 0.200$	93%
6	$\alpha_1 = 2.571$	92%
	$\alpha_2 = -2.919$	96%
	$\alpha_3 = 1.528$	99%
	$\alpha_4 = 6.012$	91%
	$\alpha_5 = -0.512$	91%
	$\pi_1 = 0.170$	94%
	$\pi_2 = 0.170$	91%
	$\pi_3 = 0.170$	92%
	$\pi_4 = 0.170$	90%
	$\pi_5 = 0.170$	92%

Bibliography

Abbi, R., El-Darzi, E., Vasilakis, C., and Millard, P. Analysis of stopping criteria for the EM algorithm in the context of patient grouping according to length of stay. In *Intelligent Systems, 2008. 4th International IEEE Conference*, pages 3–9, 2008.

Agresti, A. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002.

Agresti, A. *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, 2nd edition, 2007.

Agresti, A. *Analysis of Ordinal Categorical Data*. Wiley Series in Probability and Statistics. Wiley, 2nd edition, 2010.

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

Ahn, J., Mukherjee, B., Banerjee, M., and Cooney, K. A. Bayesian inference for the stereotype regression model: Application to a case-control study of prostate cancer. *Statistics in Medicine*, 28(25):3139–3157, 2009.

Ahn, J., Mukherjee, B. R., Gruber, S. B., and Sinha, S. Missing exposure data in stereotype regression model: application to matched case-control study with disease subclassification. *Biometrics*, 67(2):546–558, 2011.

Akaike, H. Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *2nd International Symposium on Information Theory*, pages 267–281, 1973.

Anders, R. and Batchelder, W. H. Cultural consensus theory for the ordinal data case. *Psychometrika*, Published online:1–31, 2013. URL `http://link.springer.com/article/10.1007/s11336-013-9382-9`.

Anderson, D. R., Burnham, K. P., and White, G. C. Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics*, 25(2):263–282, 1998.

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

Andrews, R. L. and Currim, I. S. A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, 40(2):235–243, 2003.

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

Banfield, J. D. and Raftery, A. E. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

Barker, R. J. and Link, W. A. Bayesian multimodel inference by RJMCMC: A Gibbs sampling approach. *The American Statistician*, 67(3):150–156, 2013.

Bartolini, I., Ciaccia, P., Ntoutsi, I., Patella, M., and Theodoridis, Y. The PANDA framework for comparing patterns. *Data and Knowledge Engineering*, 68(2):244–260, 2009.

Best, N. G., Spiegelhalter, D. J., Thomas, A., and Brayne, C. E. G. Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society, Series A*, 159(2):323–342, 1996.

Bezdek, J. C., Li, W. Q., Attikiouzel, Y., and Windham, M. A geometric approach to cluster validity for normal mixtures. *Soft Computing*, 1(4):166–179, 1997.

Biernacki, C. and Govaert, G. Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 29(2):451–457, 1997.

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

Biernacki, C., Celeux, G., and Govaert, G. An improvement of the NEC criterion for assessing the number of clusters in mixture model. *Pattern Recognition Letters*, 20(3):267–272, 1999.

Bock, H. H. *Automatische Klassifikation (Clusteranalyse)*. Vandenhoek and Ruprecht, Göttingen, 1974.

Bock, R. D. and Jones, L. V. *The measurement and prediction of judgment and choice*. Holden-Day series in psychology. Holden-Day, 1968.

Böhning, D. and Seidel, W. Editorial: recent developments in mixture models. *Computational Statistics and Data Analysis*, 41(3):349–357, 2003.

Böhning, D., Seidel, W., Alfò, M., Garel, B., Patilea, V., and Walther, G. Advances in mixture models. *Computational Statistics and Data Analysis*, 51(11):5205–5210, 2007.

Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psycometrika*, 52(3):345–370, 1987.

Bozdogan, H. *Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse Fisher information matrix*. Information and Classification (O. Opitz et al. (eds.)). Springer Berlin Heidelberg, 1993.

Bozdogan, H. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, 1(3): 69–113, 1994.

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

Brooks, S. P. Discussion to Richardson and Green (1997). *Journal of the Royal Statistical Society. Series B*, 59(4):774–775, 1997.

Brooks, S. P. Markov chain Monte Carlo method and its application. *The Statistician*, 47(1):69–100, 1998.

Brooks, S. P. and Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

Brooks, S. P. and Giudici, P. Markov chain Monte Carlo convergence assessment via two-way analysis of variances. *Journal of Computational and Graphical Statistics*, 9(2):266–285, 2000.

Brooks, S. P, Giudici, P, and Philippe, A. Nonparametric convergence assessment for MCMC model selection. *Journal of Computational and Graphical Statistics*, 12 (1):1–22, 2003.

Burnham, K. P. and Anderson, D. R. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer, 2nd edition, 2002.

Cappé, O., Robert, C., and Rydén, T. Reversible jump, birth-and-death, and more general continuous time MCMC samplers. *Journal of the Royal Statistical Society Series B*, 65(3):679–700, 2003.

Carlin, B. P. and Chib, S. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57(3):473–484, 1995.

Castelloe, J. and Zimmerman, D. Convergence assessment for reversible jump MCMC samplers. Technical Report 313, SAS Institute, Cary, North Carolina, 2002. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.759&rep=rep1&type=pdf`.

Celeux, G. Bayesian inference for mixtures: The label switching problem. In *Proceedings in Computational Statistics 1998 (COMPSTAT98)*, pages 227–232. Physica-Verlag HD, 1998.

Celeux, G. and Soromenho, G. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

Celeux, G., Hurn, M., and Robert, C. P. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. Deviance information criteria for missing data models (with discussion). *Bayesian Analysis*, 1(4): 651–674, 2006.

Chen, L.-C., Yu, P. S., and Tseng, V. S. A weighted fuzzy-based biclustering method for gene expression data. *International Journal of Data Mining and Bioinformatics*, 5(1):89–109, 2011.

Chib, S. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

Cole, S. R. and Ananth, C. V. Regression models for unconstrained partially or fully constrained continuation odds ratios. *International Journal of Epidemiology*, 30(6):1379–1382, 2001.

Dalenius, T. and Hodges, J. L. Minimum variance stratification. *Journal of the American Statistical Association*, 54(285):88–101, 1959.

DeIorio, M. and Robert, C. P. Discussion on "Bayesian measures of model complexity and fit" (by D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. van der Linde). *Journal of the Royal Statistical Society, Series B*, 64(4):629–630, 2002.

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

DeSantis, S. M., Houseman, E. A., Coull, B. A., Stemmer-Rachamimov, A., and Betensky, R. A. A penalized latent class model for ordinal data. *Biostatistics*, 9 (2):249–262, 2008.

DeSarbo, W. S., Fong, D. K. H., Liechty, J., and Kim Saxton, M. A hierarchical Bayesian procedure for two-mode cluster analysis. *Psychometrika*, 69(4):547–572, 2004.

Development Core Team, R. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Ehrgott, M. *Multicriteria Optimization*. Springer, 2006.

Engelman, L. and Hartigan, J. A. Percentage points of a test for clusters. *Journal of the American Statistical Association*, 64(328):1647–1648, 1969.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. *Cluster Analysis*. John Wiley and Sons, Ltd, Chichester, UK, 5th edition, 2011.

Feldmann, U. and König, J. Ordinal classification in medical prognosis. *Methods of information in medicine*, 41(2):154–163, 2002.

Fernández, D., Arnold, R., and Pledger, S. Mixture-based clustering for the ordered stereotype model. *Computational Statistics and Data Analysis*, 2014a. URL `http://www.sciencedirect.com/science/article/pii/S016794731400317X`.

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

Figueredo, M. A. T. and Jain, A. K. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern analysis and Machine Intelligence*, 24(3):1–16, 2002.

Fonseca, J. R. S. The application of mixture modeling and information criteria for discovering patters of coronary heart disease. *Quantitative Methods in Medical Sciences*, 3(4):292–303, 2008.

Fonseca, J. R. S. and Cardoso, M.G.M.S. Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, 11(2):155–173, 2007.

Fraley, C. and Raftery, A. E. How many clusters? Which clustering method? answers via model-based cluster analysis. *Computer Journal*, 41(8):578–588, 1998.

Fraley, C. and Raftery, A. E. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Fraley, C. and Raftery, A. E. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, 2007.

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

Frühwirth-Schnatter, S. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 453(96):194–209, 2001.

Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*. Wiley and Sons, New York, 2006.

Gamerman, D. and Lopes, H. F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science. Taylor & Francis, 2006.

Gelman, A. and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Geman, S. and Geman, D. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

Geweke, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4 (J.M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.))*, pages 169–194, 1992.

Geyer, C. J. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC, 2011.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC Interdisciplinary Statistics Series. Chapman and Hall, 1996.

Goodman, L. A. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552, 1979.

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

Govaert, G. and Nadif, M. An EM algorithm for the block mixture model. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 27(4):643–647, 2005.

Govaert, G. and Nadif, M. Latent block model for contingency table. *Communications in Statistics - Theory and Methods*, 39(3):416–425, 2010.

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

Green, P. J. and Hastie, D. I. Reversible jump MCMC. *Genetics*, 153(3):1391–1403, 2009.

Greenland, S. Alternative models for ordinal logistic regression. *Statistics in Medicine*, 13(16):1665–1677, 1994.

Haefner, J. W. *Modeling Biological Systems: Principles and Applications*. Modeling Biological Systems: Principles and Applications. Springer, 2005.

Hammond, R. K. and Bickel, J. E. Reexamining discrete approximations to continuous distributions. *Decision Analysis*, 10(1):6–25, 2013.

Hartigan, J. A. and Kleiner, B. Mosaics for contingency tables. In *Proceedings of the 13th Symposium on the Interface between Computer Sciencies and Statistics*, pages 268–273. Springer-Verlag, 1981.

Heidelberger, P. and Welch, P. D. Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1145, 1983.

Heilbron, D. Generalized linear models for altered zero probabilities and overdispersion on count data. Technical report, Department of Epidemiology and Biostatistics, University of California, San Francisco, 1989.

Hennig, C. and Liao, T. F. How to find an appropriate clustering for mixed type variables with application to socioeconomic stratification. *Journal of the Royal Statistical Science, Series C (Applied Statistics)*, 62(3):309–369, 2013.

Hilbe, J. M. *Negative Binomial Regression*. Cambridge University Press. New York, 2008.

Hjorth, J. S. U. *Computer intensive statistical methods: validation, model selection and bootstrap*. Chapman and Hall, London, 1994.

Hoffman, D. L. and Franke, G. R. Correspondence analysis: Graphical representation of categorical data in marketing research. *Journal of Marketing Research*, 23 (3):213–227, 1986.

Holtbrugge, W. and Schumacher, M. A comparison of regression models for the analysis of ordered categorical data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40(2):249–259, 1991.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985.

Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D., and Warton, D. I. Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, 2014.

Hurn, M., Justel, A., and Robert, C. P. Estimating mixture of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.

Hurvich, C. M. and Tsai, C. L. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

Hyndman, R. J. and Fan, Y. Sample quantiles in statistical packages. *Statistical Computing*, 50(4):361–365, 1996.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2014.

Jasra, A., Holmes, C. C., and Stephens, D. A. MCMC and the label switching problem in Bayesian mixture models. *Statistical Science*, 20(1):50–67, 2005.

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

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.

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

Keribin, C. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 6:49–66, 2000.

Keribin, C., Brault, V., Celeux, G., and Govart, G. Estimation and selection for the latent block model on categorical data. Technical report, INRIA research report, 2012.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

Koehler, K. J. and Gan, F. F. Chi-squared goodness-of-fit tests: Cell selection and power. *Communications in Statistics - Simulation and Computation*, 19(4):1265–1278, 1990.

Kotsiantis, S. and Kanellopoulos, D. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58, 2006.

Kraskov, A., Stögbauer, H., Andrzejak, R. G., and Grassberger, P. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70(2):278–284, 2005.

Kullback, S. and Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Kuss, O. On the estimation of the stereotype regression model. *Computational Statistics and Data Analysis*, 50(8):1877–1890, 2006.

Labiod, L. and Nadif, M. Co-clustering for binary and categorical data with maximum modularity. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 1140–1145, 2011.

Lanning, J. M. and Bozdogan, H. *Statistical Data Mining and Knowledge Discovery*. Statistics: Computer science. CRC Press, 2004.

Lavallee, P. and Hidiriglou, M.A. On the stratification of skewed populations. *Survey Methodology*, 14(1):33–43, 1988.

Lee, J. and Wong, D. W. S. *Statistical Analysis with ArcView GIS*. Wiley, 2001.

Lee, J. A. and Verleysen, M. *Nonlinear Dimensionality Reduction*. Springer, 2007.

Lee, K., Marin, J. M., Robert, C., and Mengersen, K. Bayesian inference on mixtures of distributions. In *Proceedings of the Platinum Jubilee of the Indian Statistical Institute 776*, 2008.

Leroux, B. G. Consistent estimation of a mixing distribution. *Annals of Statistics*, 20(3):1350–1360, 1992.

Lewis, S. J. G., Foltynie, T., Blackwell, A. D., Robbins, T. W., Owen, A. M., and Barker, R. A. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery and Psychiatry*, 76(3):343–348, 2003.

Link, W. A. and Barker, R. J. *Bayesian inference with ecological applications*. Academic Press, London, UK, 2010.

Link, W. A. and Eaton, M. J. On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3(1):112–115, 2012.

Liu, I. and Agresti, A. The analysis of ordered categorical data: An overview and a survey of recent developments. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 14(1):1–73, 2005.

Lunt, M. Stereotype ordinal regression. *Stata Technical Bulletin*, 10(61), 2001.

Lunt, M. Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels. *Statistics in Medicine*, 24(9): 1357–69, 2005.

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

Manly, B. F. J. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. London: Chapman & Hall, 3rd edition, 2007.

Marin, J. M. and Robert, C. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer Texts in Statistics. Springer, 2007.

Marin, J. M., Mengersen, K., and Robert, C. *Bayesian Modelling and Inferences on Mixtures of Distributions*. Handbook of Statistics, volume 25 (eds. D. Dey and C. R. Rao). Springer-Verlag, New York, 2005.

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

Maxwell, A. E. *Analysing Qualitative Data*. Methuen, New York, 1961.

McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society*, 42(2):109–142, 1980.

McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. London: Chapman & Hall, 2nd edition, 1989.

McCune, B. and Grace, J. B. *Analysis of Ecological Communities*, volume 28. MjM Software Design, 2002.

McLachlan, G. and Ng, S. K. A comparison of some information criteria for the number of components in a mixture model. Technical report, Brisbane: Department of Mathematics, University of Queensland, 2000.

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

McLachlan, G. J. The classification and mixture maximum likelihood approaches to cluster analysis. *Handbook of Statistics*, 2(299):199–208, 1982.

McLachlan, G. J. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36(3):318–324, 1987.

McLachlan, G. J. and Basford, K. E. *Mixture models: inference and applications to clustering*. Statistics, textbooks and monographs. M. Dekker, 1988.

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

McPartland, D. and Gormley, I. C. *Clustering Ordinal Data via Latent Variable Models*. Algorithms from and for Nature and Life. Studies in Classification,

Data Analysis, and Knowledge Organization. Springer International Publishing, 2013.

McQuarrie, A., Shumway, R., and Tsai, C.-L. The model selection criterion AICu. *Statistics and Probability Letters*, 34(3):285–292, 1997.

Meila, M. Comparing clusterings: an axiomatic view. In *ICML 2005: Proceedings of the 22nd International Conference on Machine Learning*, pages 577–584. ACM Press, 2005.

Meila, M. Comparing clusterings-an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

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

Mengersen, K. L. and Robert, C. P. Testing for mixtures: a Bayesian entropic approach. *In Bayesian Statistics 5 (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith), Oxford University Press*, 4(9):255–276, 1996.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Moustaki, I. A latent variable model for ordinal variables. *Applied Psychological Measurement*, 24(3):211–233, 2000.

Mukherjee, B., Ahn, J., Liu, I., Rathouz, P. J., and Sanchez, B. Fitting stratified proportional odds models by amalgamating conditional likelihoods. *Statistics in Medicine*, 27(24):4950–4971, 2008.

Mullahy, J. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

Nobile, A. *Bayesian Analysis of Finite Mixture Distributions*. PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, 1994. URL `http://www.stats.gla.ac.uk/~agostino`.

Ntoutsi, I., Pelekis, N., and Theodoridis, Y. *Pattern Comparison in Data Mining: a Survey*. Research and Trends in Data Mining Technologies and Applications (ed. David Taniar). Idea Group Pub., 2006.

Phillips, D. B. and Smith, A.F.M. *Bayesian model comparison via jump diffusions*. In Markov chain Monte Carlo in Practice (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). Chapman and Hall, London, 1996.

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

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

Plummer, M. Penalized loss functions for Bayesian model comparison. *Biostatistics*, 9(3):523–539, 2008.

Plummer, M., Best, N. G., Cowles, K., and Vines, K. Coda: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006. URL `http://CRAN.R-project.org/doc/Rnews/Rnews_2006-1.pdf`.

Plummer, M., Best, N. G., Cowles, K., Vines, K., and Sarkar, D. *Package Coda (Version 0.16-1): Output analysis and diagnostics for MCMC*. Available at: http://cran.r-project.org/web/packages/coda/, June 2012.

Preedalikit, K. *Joint Modeling of Longitudinal Ordinal Data on Quality of Life and Survival*. PhD thesis, MSOR School-Victoria University of Wellington, 2012. URL `http://hdl.handle.net/10063/2543`.

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

Raftery, A. E. *Hypothesis testing and model selection. In Markov chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Ray, S. and Ren, D. On the upper bound of the number of modes of a multivariate normal mixture. *Journal of Multivariate Analysis*, 108:41–52, 2012.

Richardson, S. and Green, P. J. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B*, 59 (4):731–792, 1997.

Robert, C. and Casella, G. *Introducing Monte Carlo Methods with R*. Springer Science and Business Media, 2010.

Roberts, F. and Tesman, B. *Applied Combinatorics (2nd ed.)*. CRC Press, 2011.

Rocci, R. and Vichi, M. Two-mode multi-partitioning. *Computational Statistics and Data Analysis*, 52(4):1984–2003, 2008.

Roeder, K. and Wasserman, L. Practical Bayesian density estimation using mixture of normals. *Journal of the American statistical Association*, 92(439):894–902, 1997.

Rogers, A. *Statistical Analysis of Spatial Dispersion: The Quadrat Method*. Monographs in Spatial and Environmental Systems Analysis. Pion, 1974.

Saei, A., Ward, J., and McGilchrist, C. A. Threshold models in a methadone programme evaluation. *Statistics in Medicine*, 15(20):2253–2260, 1996.

Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

Self, S. G. and Liang, K.-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

Siddhartha, C. and Greenberg, E. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

Sisson, S. A. and Fan, Y. A distance-based diagnostic for trans-dimensional Markov chains. *Statistics and Computing*, 17(4):357–367, 2007.

Skrondal, A. and Rabe-Hesketh, S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Taylor & Francis, 2004. ISBN 9780203489437.

Snell, E. J. A scaling procedure for ordered categorical data. *Biometrics*, 20(3): 592–607, 1964.

Soromenho, G. Comparing approaches for testing the number of components in a finite mixture model. *Computational Statistics*, 9(1):65–78, 1994.

Sperrin, M., Jaki, T., and Wit, E. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20(3): 357–366, 2010.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64(4):583–639, 2002.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, 76(3):485–493, 2014.

Stahl, D. and Sallis, H. Model-based cluster analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(4):341–358, 2012.

Steele, R. J. and Raftery, A. E. *Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models*. In Frontiers of Statistical Decision Making and Bayesian Analysis. Springer, 2010.

Stephens, M. Dealing with label switching in mixture models. *Journal of Royal Statistical Society, Series B*, 62(4):795–809, 2000a.

Stephens, M. Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74, 2000b.

Stevens, S. S. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

Strehl, A. and Ghosh, J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(1):583–617, 2002.

Tobin, J. Estimation for relationships with limited dependent variables. *Econometrica*, 26(1):24–36, 1958.

Tozer, M. G. and Bradstock, R. A. Fire-mediated effects of overstorey on plant species diversity and abundance in an eastern Australian heath. *Plant Ecology*, 164(2):213–223, 2002.

Umbach, D. M. and Wilcox, A. J. A technique for measuring epidemiologically useful features of birthweight distributions. *Statistics in Medicine*, 15(13):1333–1348, 1998.

Van der Aart, P. and Smeenk-Enserink, N. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology*, 25(1):1–45, 1974.

Vermunt, J. K. The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement*, 25(3):283–294, 2001.

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

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(1):2837–2854, 2010.

Wagenmakers, E. J., Lee, M., Lodewyckx, T., and Iverson, G. J. *Bayesian Versus Frequentist Inference*. Springer, 2008.

Warton, D. I., Wright, S. T., and Wang, Y. Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, 3(1): 89–101, 2012.

Westhoff, V. and van der Maarel, E. The Braun-Blanquet approach. In H., Whittaker R., editor, *Classification of Plant Communities*, pages 287–328. Junk, The Hague, 1978.

Whittaker, R. H. Vegetation of the Great Smoky mountains. *Ecological Monographs*, 26(1):1–80, 1956. ISSN 00129615.

Wilks, S. S. The large sample distribution of the likelihood ratio for testing composite hypothesis. *Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Wismüller, A., Verleysen, M., Aupetit, M., and Lee, J. A. Recent advances in nonlinear dimensionality reduction, manifold and topological learning. In

European Symposium on Artificial Neural Networks(ESANN), 2010. URL `http://perso.uclouvain.be/michel.verleysen/papers/esann10aw.pdf`.

Wu, H.-M., Tzeng, S., and Chen, C.-H. Matrix visualization. *Handbook of Data Visualization*, pages 681–708, 2007.

Wu, X., Kumar, V., Quinlan, J. R, Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

Yee, T. W. The VGAM package. *R News*, 8(2):28–39, October 2008. URL `http://CRAN.R-project.org/doc/Rnews/`.

Yee, T. W. and Hastie, T. J. Reduced-rank vector generalized linear models. *Statistical Modelling*, 3(1):15–41, 2003.

Zhou, H. and Lange, K. L. On the bumpy road to the dominant mode. *Scandinavian Journal of Statistics*, 37(4):612–631, 2010.

Žiberna, A., Kejžar, N., and Golob, P. A comparison of different approaches to hierarchical clustering of ordinal data. *Metodološki Zvezki - Advances in Methodology and Statistics*, 1(1):57–73, 2004.