Clustering repeated ordinal data: Model based approaches using finite mixtures

by

Roy Ken Costilla Monteagudo

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
2017

Abstract

Model-based approaches to cluster continuous and cross-sectional data are abundant and well established. In contrast to that, equivalent approaches for repeated ordinal data are less common and an active area of research. In this dissertation, we propose several models to cluster repeated ordinal data using finite mixtures. In doing so, we explore several ways of incorporating the correlation due to the repeated measurements while taking into account the ordinal nature of the data.

In particular, we extend the Proportional Odds model to incorporate latent random effects and latent transitional terms. These two ways of incorporating the correlation are also known as parameter and data dependent models in the time-series literature. In contrast to most of the existing literature, our aim is classification and not parameter estimation. This is, to provide flexible and parsimonious ways to estimate latent populations and classification probabilities for repeated ordinal data.

We estimate the models using Frequentist (Expectation-Maximization algorithm) and Bayesian (Markov Chain Monte Carlo) inference methods and compare advantages and disadvantages of both approaches with simulated and real datasets. In order to compare models, we use several information criteria: AIC, BIC, DIC and WAIC, as well as a Bayesian Non-Parametric approach (Dirichlet Process Mixtures). With regards to the applications, we illustrate the models using self-reported health status in Australia (poor to excellent), life satisfaction in New Zealand (completely agree to completely disagree) and agreement with a reference genome of infant gut bacteria (equal, segregating and variant) from baby stool samples.

Acknowledgments

To write a PhD dissertation is like building a house from scratch. You need to draw the plans, choose the land, build the foundations and so on. It is a long and complex process but also a hugely rewarding one, once you see the house completed. Of course, you need the help of others to carry out such a mammoth endeavour.

Many thanks to my advisors Ivy Liu and Richard Arnold for their unconditional support over nearly four years. I would have not been able to do it without you. Sincere thanks also to Shirley Pledger, Daniel Fernández, Petros Hadjicostas, Patricia Huambachano, and Peter Donelan from the School of Mathematics and Statistics for your advice, encouragment and friendship in this journey.

I am also very grateful to my informal advisors: Peter Müller from University of Texas-Austin, Joseph Bulbulia from Victoria University of Wellington, Dean Hyslop from Motu and Russell Millar from University of Auckland. I feel very fortunate to have had your advice and mentorship over the different stages of my PhD. A special mention to Miles Benton from Queensland University of Technology and Aaron Darling from University of Technology Sydney for our conversations about Genomics (and baby poo). Last but not least, I would like to acknowledge John Hinde, Lynn Hunt and Laura Dumitrescu who reviewed this dissertation and provided many insightful comments and words of encouragment. Needless to say, all remaining errors and omissions are my responsability.

I want to dedicate my doctoral dissertation to my family: my mum Lucia, my dad Valerio, and my brother Ronny in Peru; my partner Emma, and our son Amaru in New Zealand; and our son Jesús Valentino who is flying around everywhere.

Contents

Chapter 1

Literature Review

1.1 Introduction

A variable with an ordered categorical scale is called *ordinal*. That is, ordinal data is categorical data where outcome levels have a logical order and thus its order matters. Examples of ordinal responses are: socio-economic status (low, medium, high), educational attainment (high school, vocational, undergraduate, postgraduate), disease severity (not infected, initial, medium, advanced), health status (poor, fair, good, excellent), agreement with a given statement (strongly disagree, disagree, neutral, agree, strongly agree) and any other variables that use the Likert scale. Conversely, a categorical variable with an unordered scale is called *nominal*. In this case, categories differ in quality not in quantity (Agresti 2013). Religious affiliation (Non-religious, Christian, Muslim, Jewish, Buddhist, Other), geographical location (North, East, South, West), preferred method of commuting (bus, train, bike, walk, other) amongst others are examples of nominal variables.

Analyses of ordinal data are very common but often do not fully exploit their ordinal nature. First, ordinal outcomes are treated as continuous by assigning numerical scores to ordinal categories. Doing this equates to assuming that the categories are equally spaced in the ordinal scale which

might be an unnecessary and restrictive assumption. Secondly, methods for the identification of latent groups, patterns, and clusters in ordinal data lag behind equivalent approaches for continuous, binary, nominal and count data. In particular, traditional clustering approaches such hierarchical clustering, association analysis, and partition optimization methods like k-means clustering; are not based on likelihoods and thus statistical inference tools are not available. For instance, model selection criteria can not be used to evaluate and compare different models. Thirdly, another common approach is to ignore the order of the categories altogether and thus treat the data as nominal. By ignoring the ranked nature of the categories this approach reduces its statistical power for inference.

Further challenges are posed when repeated measurements of an ordinal response are made for each unit, such as in longitudinal studies. For these two-way data (unit by time period), the correlation structure among repeated measures needs also to be accounted for. The correlation structure could be generalised to the analysis of three-way data where for each unit there are several ordinal responses at a given moment and these are repeated over time (unit by question by time period). Moreover, two-way data could also be combined with observations of an additional response variable giving rise to joint models where a common latent variable explains both the repeated ordinal and the additional outcomes. For instance, consider as a motivating example the health status of person measured several times, and whether or not they were welfare beneficiaries over that period. Both in turn could depend on a latent variable such as deprivation. The research goal could be to group individuals in order to identify those at the greatest risk of living on the benefit and ultimately estimate their latent deprivation level. Resulting models are thus markedly more complex both in mathematical and computational terms. We now present a review of the existing models in the literature.

1.2 Models for Ordinal Data

Ordinal data is often analysed by modelling the cumulative probabilities of the ordinal response and using a link function, usually logit or probit. Although methods for categorical data started off in the 1960s, Snell (1964), Bock & Jones (1968), models for ordinal data were mostly developed after the influential articles by McCullagh (1980) on modelling of cumulative probabilities using a logit link and Goodman (1979) on loglinear models for odds ratios of ordered categories. Substantial developments have been made since then and are well documented elsewhere by Liu & Agresti (2005) and Agresti (2013). Here we will review in detail the most relevant models for our purposes and briefly mention the rest.

Cumulative logit models

The Proportional Odds model

The Proportional Odds Model (POM) by McCullagh (1980) is a cumulative logit model and is the most popular model for analysing ordinal data. It links the logits of the cumulative probabilities with a set of predictors. For a ordinal response Y with q ordered categories and a set of predictors $x = (x_1, \ldots, x_m)'$ the model can be written as

$$\text{Logit}[P(Y \leq k|x)] = \mu_k - \beta' x \qquad k = 1, \ldots, q-1$$

where $\mu_1 < \mu_2 \cdots < \mu_{q-1}$. These parameters μ_k are called cut points but also regarded as nuisance parameters because they are often of no or little interest. This model has $q-1$ equations, that is it applies simultaneously to all $q-1$ cumulative logits. The parameter β captures the effect of the predictors on the cumulative probabilities and is the same for all levels of the cumulative probability (β is the same for all k). This *Proportional Odds* property gives the model its name and implies that the odds ratios for describing effects of explanatory variables on the ordinal response are the

same for each of the possible ways of collapsing the q ordinal categories to a binary variable.

We use a parametrisation with a negative sign, because it allows the coefficients β to have the usual directional meaning of the predictor on the response. That is, for predictor m, $\beta_m \geq 0$ implies that Y is more likely to fall at the high end of the ordinal scale.

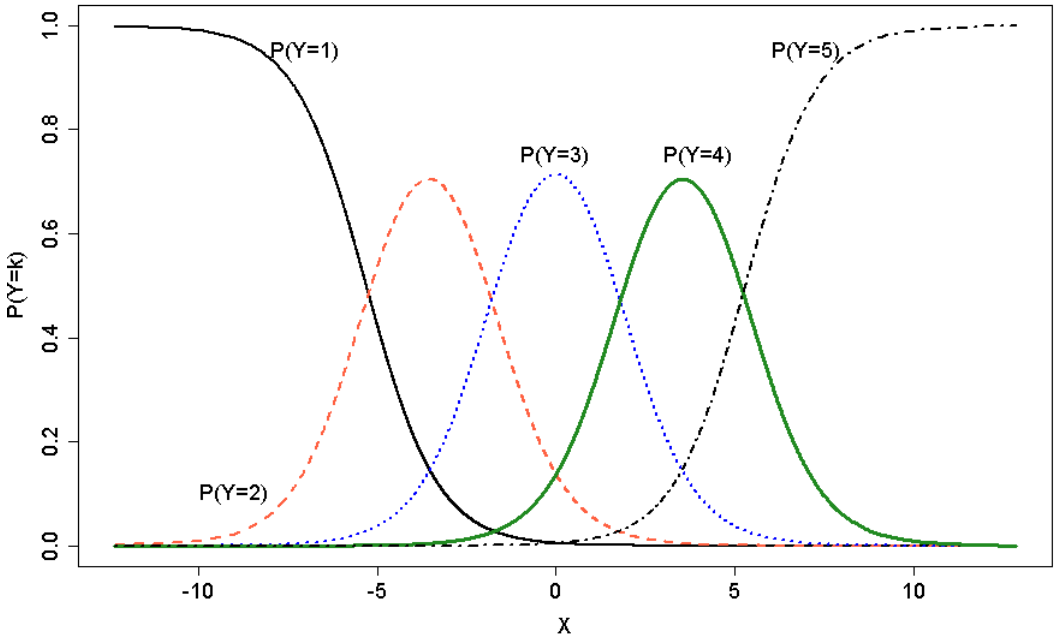


Figure 1.1: Individual category probabilities for the POM with five response categories

Figure 1.1 shows a graphical representation of the POM for five response categories and one continuous predictor. As it can be seen, $P(Y = k)$ has the same shape for all the ordinal categories ($k = 2, \ldots, q-1$) and differs only in its location.

Alternatively, the POM has also a latent variable representation (Anderson & Philips 1981). Assuming that the ordinal response Y comes from

an underlying continuous response Y^* which follows a standard logistic distribution conditional on x such that $Y = k$ if $\mu_{k-1} \leq Y^* \leq \mu_k$, then the POM holds for Y. In other words, the POM could also be represented as $Y^* = \beta' x + \epsilon$ where $\epsilon \sim$ Logistic$(0, \pi^2/3)$. Figure 1.2 shows a graphical representation, the ordinal response Y (right Y-axis) falls in category $k = 1, 2, 3, 4$ when the unobserved continuous response Y^* falls in the k^{th} interval of values. The slope of the regression line is β.

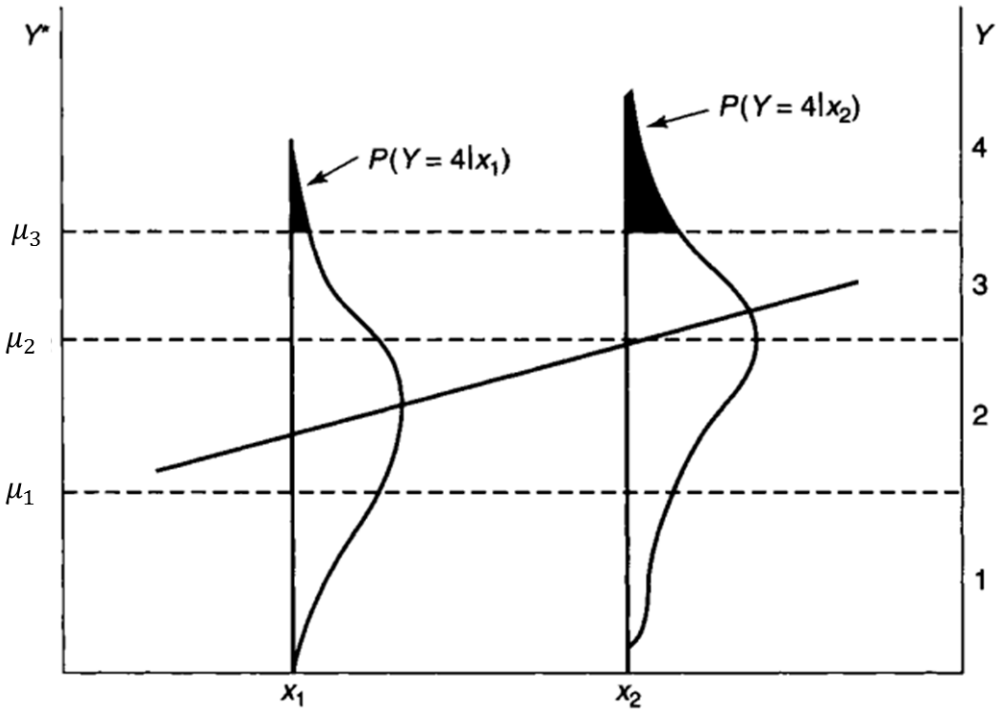


Figure 1.2: Latent variable representation of the POM, reprinted from Agresti (2013).

Maximum Likelihood (ML) methods are often used to fit cumulative logit models. ML estimates of the model parameters are obtained using iterative methods that solve the likelihood equations for all the cumulative logits, e.g. $q - 1$ equations in the case of the POM described above. Walker & Duncan (1967), McCullagh (1980) proposed the Fisher scoring

algorithm, an iteratively reweighted least squares algorithm, for this task. A sufficiently large n guarantees a global maximum but a finite n does not. In the latter, the ML function may exhibit local maxima or not have one at all (McCullagh 1980).

When the POM fits poorly or the proportional odds assumption is inadequate, Liu & Agresti (2005) proposed the following potential alternative strategies. (i) Trying a model with separate effects, β_k instead of β. This model however places additional constraints in the set of parameters μ and β to make sure that the cumulative probabilities are non-decreasing; (ii) Trying different link functions, (iii) Adding interactions or in general additional terms to the linear predictor; (iv) Adding dispersion terms; (v) Allowing separate effects, like in (i), for some but not all predictors. This model, introduced by Peterson & Harrell (1990), is called the *Partial Proportional Odds* model (PPOM); (vi) Using a model for nominal responses, e.g. Baseline logit. We next focus on option (iii) as it the most relevant for the purposes of this proposal. For a complete treatment see Liu & Agresti (2005).

There are several ways to include additional terms to the linear predictor when there is lack of proportional odds. Here we present the Trend Odds Model (TOM) by Capuano & Dawson (2012) that will be extented later to the clustering case. The TOM is a monotone constrained non-proportional odds models that uses a logit link for the cumulative probability and adds an extra parameter γ to the linear predictor. Setting an arbitrary scalar t_k that varies by ordinal outcome (k), the TOM has the form

$$\text{Logit}[P(Y \leq k|x)] = \mu_k - (\beta + \gamma t_k)'x \qquad t_k \leq t_{k+1}; k = 1, \ldots, q-1$$

where $\mu_k - \mu_{k-1} \geq \gamma(t_k - t_{k-1})x, \forall x$ is an additional constraint required to make sure the cumulative probabilities are non-decreasing. Intrinsically, therefore the TOM is a constrained model where for a given value of the

predictor, the odds parameter increases or decreases in a monotonic manner (γt_k with $t_k \leq t_{k+1}$) across the ordinal outcomes (k). Figure 1.3 shows a graphical representation of the TOM for five response categories and one continuous predictor. In contrast to the POM, figure 1.1, the probabilities for the ordinal responses $P(Y = k)$ differ not only in their location but also in shape. For instance, the probability that the response is equal to the second category $P(Y = 2)$ is no longer symmetric. Similarly, the probabilities that the response is equal to the first ($P(k = 1)$) and last ($P(k = 5)$) categories are no longer mirror images of each other.

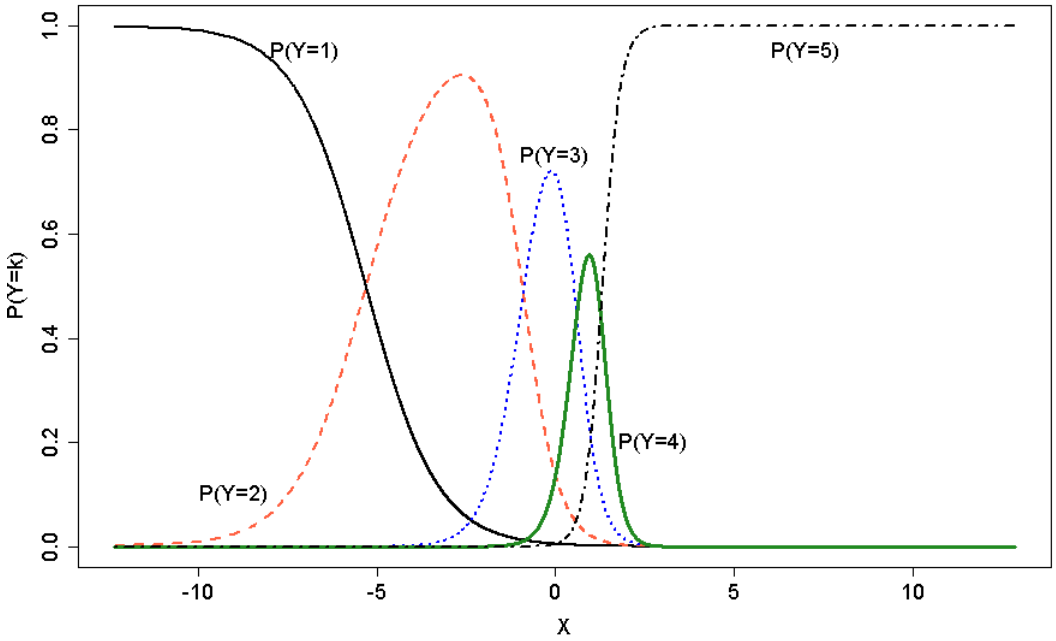


Figure 1.3: Individual category probabilities for the TOM with five response categories

Capuano & Dawson (2012) showed that the TOM is related to logistic, normal and exponential underlying latent variables and belongs to the class of constraint non-proportional odds models by Peterson & Harrell

(1990).

Other multinomial models

Alternative probability models to analyse ordinal data include: cumulative link models, continuation-ratio logit models and adjacent-categories logit model. An important related model is the Stereotype model by Anderson (1984). Nested between the adjacent-categories logit model with proportional odds and the general baseline-category logit model, it captures any potential lack of proportionality by introducing new parameters for each category. Fernández et al. (2016) extend the Stereotype model to perform model based cluster analysis, see section 1.4. We note that these models belong within the class of multivariate generalised linear models (Multivariate GLM) whenever the response has a distribution in the exponential family (McCullagh 1980, Thompson & Baker 1981, Fahrmeir & Tutz 2001).

1.3 Models for repeated ordinal data

Repeated ordinal data arise when an ordinal response is recorded at various occasions for each subject or unit, such as in longitudinal studies. We next discuss three main approaches to analyse such data: marginal models, subject-specific models and transitional models (Diggle et al. 2002, Vermunt & Hagenaars 2004, Agresti 2013).

Marginal models, also known as population-averaged models, capture the effect of the predictor averaged over all the observations. Assume for simplicity that all responses repeat the same number of times T and let Y_{it} be an ordinal response with q categories for individual i in occasion t. The marginal model with cumulative logit link has the form

$$\text{Logit}[P(Y_{it} \leq k|x_{it})] = \mu_k - \beta' x_{it} \qquad k = 1, \ldots, q-1; i = 1, \ldots, n; t = 1, \ldots, T.$$

where $x_{it} = (x_{it1}, x_{it2}, \ldots, x_{itm})'$ contains the values of the m predictors for individual i at occasion t. Model fitting is mostly performed using a generalised estimating equations (GEE) approach. This approach is a quasi likelihood method that only specifies the marginal regression models, over individuals i as in the equation above, and a working correlation structure, a guess specified by the analyst, among the T responses. Lipsitz et al. (1994) and Toledano & Gatsonis (1996) presented cumulative logit and probit models for repeated ordinal responses.

Marginal models focus on the marginal distribution of Y by averaging over individual responses and treat the joint dependence structure as nuisance. Given that our aim is to explicitly classify subjects into latent clusters, we will not be using population-averaged approach in this dissertation.

In contrast to that, subject-specific models describe effects at the individual or unit level. They are known by many names in the literature: conditional models, mixed effects models, random-effects models, and multilevel models. They jointly model the distribution of the response and the individual effects. Random effects models belong to the class of generalised linear mixed models (GLMM) when the response has a distribution in the exponential family. Individual effects are assumed to follow a certain probability distribution and hence their name of random effects. In our case, we use the random effects to capture the dependence among repeated responses but they could more generally be used to capture subject heterogeneity, unobserved covariates and other forms of overdispersion. The cumulative logit with random effects by subject has the form

$$\text{Logit}[P(Y_{it} \leq k|x_{it})] = \mu_k - \beta' x_{it} - a_i$$

where $k = 1, \ldots, q-1;\ i = 1, \ldots, N;\ t = 1, \ldots, T$ and $a_i \sim N(0, \sigma^2)$. This is the simplest model and is also known as the random intercept model. It could be extended to other ordinal models using different link functions as well as continuation-ratio logit models. ML estimation of these mod-

els is based on the marginal likelihood that integrates out the random effects. For simple cases like a random intercept Gauss-Hermite quadrature is used.

In general, multiple random effects are possible but fitting for more than two terms is challenging (Tutz & Hennevogl 1996, McCulloch et al. 2008) due to the fact that the dimensionality of the integrals that need to be solved numerically grows with the number of random effects. Higher-dimensional integrals are approximated through Monte Carlo simulation or pseudo-likelihood methods such as adaptative quadrature (Naylor & Smith 1982, Skrondal & Rabe-Hesketh 2004) and sequential Gaussian quadrature (Heiss 2008, Bartolucci et al. 2014). Of note here is the remark made by Pinheiro & Bates (1995) that quadrature and adaptive quadrature are deterministic versions of Monte Carlo integration and importance sampling.

Quadrature methods need the selection of an adequate number of integration points. This is, however, a non- straightforward task and it has to be done case by case depending on the data at hand. For instance, in fitting a mixture of latent autoregressive models, Bartolucci et al. (2014) started with 21 quadrature points for a given number of mixture components and then increased it by 10 until convergence of the estimated log-likelihood was achieved. This scheme lead them to use 51 and 61 quadrature points for mixtures with one to 4 components. Moreover, as in all Frequentist estimation, the above methods only provide point estimates for the model parameters and confidence intervals need to be estimated in a separate step, usually using the Fisher information matrix which also needs to be approximated.

In contrast to that, Bayesian approaches provide an attractive way forward. From the outset, Bayesian estimation aims to simulate the posterior distribution of parameters conditional on the data and thus provides estimates of the model parameters and their related uncertainty. Secondly, although multiple random effects might be more complex and take longer to simulate, thanks to the theory of Markov chains we are sure that a well

constructed MCMC chain will be able to efficiently explore any target distribution in finite time. Of course, a Bayesian approach is not a silver bullet that could be used for free. Issues such as the selection of priors, assessment of the convergence of the MCMC chain, and the design of MCMC moves for efficient exploration of the target distribution are of uttermost importance when using a Bayesian approach. Thanks to the advances in the Bayesian literature in the last decades, see for example Johnson & Albert (1999), Robert & Casella (2005), Frühwirth-Schnatter (2006), Gelman et al. (2014), Müller et al. (2015), we now have a standard set of tools to tackle these issues. This together with the increase of computational power make Bayesian approaches a natural choice in complex models like the above. Chapter 5 presents a finite mixture for repeated ordinal data with latent random effects that follow a random walk with cluster-specific variance, and estimates it within a Bayesian framework.

Finally, transitional models include also past responses as predictors. That is, they model the ordinal response Y_t conditional on past responses $Y_{t-1}, Y_{t-2}, \ldots$ and other explanatory variables x_t. A very popular transitional model is the first-order Markov model in which Y_t is assumed to depend only on Y_{t-1} and covariates of time t. For example, Kedem & Fokianos (2005) used a cumulative logit transitional model in the context of a longitudinal medical study. In our case, Chapter 6 presents a finite mixture for repeated ordinal data with transitional terms by latent cluster and occasion and estimates it within a Bayesian framework.

As remarked by Liu & Agresti (2005), the use of any of these three approaches depends on the problem at hand. That is, an approach should be chosen according to whether interpretations are needed at the population level, subject-specific predictions are of relevance, or whether or not it is important to describe effects of explanatory variables conditional on past responses. Furthermore, estimated effects can have different magnitude depending on the approach taken. For example, in transitional models the interpretation and magnitude of the effect of the past responses on the or-

dinal response depends on how many previous observations are include in the model. Also the effects of the other explanatory variables diminish markedly (Agresti 2013). In addition to that, effects in a subject-specific model are larger in magnitude than those in a population-averaged model.

1.4 Model based clustering analysis for ordinal data

Traditional cluster analysis approaches treat ordinal responses as continuous and reduce the dimensionality of the data by using the eigenvalues and matrix decomposition. Amongst others, hierarchical clustering (Kaufman & Rousseeuw 1990) , association analysis (Manly 2005), and partition optimization methods like k-means clustering (Lewis et al. 2003), follow this approach. Since these approaches are not based on likelihoods, statistical inference tools are not available and model selection criteria can not be used to evaluate and compare different models.

Model based approaches, such as Kendall's τ_b (Kendall 1945), Goodman-Kruskal's γ (Goodman & Kruskal 1954) and Somers' d (Somers 1962); also exists. However, they use distance metrics and similarity measures and thus do not fully exploit the ordinal structure of the data. Their associated statistical tests rely on Monte Carlo methods, testing only the sample at hand and not more general hypothesis about the data generating process. Hastie et al. (2009) presents full details.

Model based clustering using finite mixtures have been proposed by several authors (Everitt & Hand 1981, McLachlan & Peel 2000). See a recent literature review by Fernández et al. (2017). This approach poses probabilistic models using finite mixtures which are mostly fitted using the Expectation-Maximisation algorithm (EM) (Dempster et al. 1977) and focus on either continuous, discrete or nominal responses. A major advantage of this approach is the availability of likelihoods, for the probability

models, and therefore access to various model selection criteria to evaluate and compare different models.

Finite mixtures are also known as Latent Class (LC) models in the literature of Latent Variable models (Bartholomew et al. 2011, Wedel & DeSarbo 1995). Used firstly in sociology (Lazarsfeld 1950), LC have been widely used to cluster ordinal responses, as well as continuous and categorical data. Although slight differences could be argued, McLachlan & Peel (2000) for instance points out that mixture components in LC models often represent actual underlying classes that may have a meaningful physical representation, both denominations have been used interchangeably in the literature even in early applications, see for example Aitkin's seminal paper (Aitkin et al. 1981). In addition to that, this literature makes also an interesting connection of finite mixtures with random effects models. Here mixtures are a way to estimate models with discrete random effects since the distribution of the random effects is assumed to be multinomial across the latent classes. They accordingly use the terms non-parametric random-effects and non-parametric maximum likelihood (NPML) for the model and its estimation approach (Wedel & DeSarbo 1994, Aitkin 1996, Aitkin & Alfó 1998, Vermunt & Van Dijk 2001, Alfò et al. 2016).

Also of note in the latent variable literature are Skrondal & Rabe-Hesketh (2004) and Vermunt & Magidson (2000) who implemented this approach and made them available in standard software, GLLAMM (Generalised Linear Latent Mixed Models) and Latent Gold, respectively. It is important to stress that most of this literature has focused on parameter estimation and not clustering nor classification, that is the assignment of subjects to the latent classes. On the other hand, the latent variable literature has also had an almost exclusive reliance on AIC and BIC for model comparison (Nylund et al. 2007) which might not be appropriate for finite mixtures as it tends to overestimate the number of latent clusters (McLachlan & Peel 2000).

Simultaneous clustering of row and columns is called biclustering, block

clustering or two-mode clustering. Biclustering models for binary, count and categorical data have been proposed by Biernacki et al. (2000), Pledger (2000), Govaert & Nadif (2008), Arnold et al. (2010), Labiod & Nadif (2011), Pledger & Arnold (2014). More recently, Matechou et al. (2016) and Fernández et al. (2016) have extended these models to ordinal responses. The former used the proportional odds (McCullagh 1980) and the latter the Stereotype model (Anderson 1984) and enable them to handle row, column and biclustered data. The overall aim of this dissertation is to extent these finite mixture models to the case of repeated ordinal data using the proportional odds formulation. This is done in Chapters 5, 6 and 7.

1.5 Model based clustering for repeated ordinal data

In analogy to the literature for longitudinal data, there are two main approaches for finite mixture-based clustering for repeated ordinal data: mixtures of random effects models and mixtures of transitional models which we denote here *parameter dependent* (chapter 5) *data dependent* (chapter 6) models. These names are however just conventions as both approaches in turn could be viewed as special cases of non-linear state space models for longitudinal data (Fahrmeir & Tutz 2001) and models that combined both approaches could also be meaningful and have been proposed in the literature (Bates & Neyman 1952, Heckman 1981*a*, Skrondal & Rabe-Hesketh 2014).

1.5.1 Parameter dependent models

Parameter dependent models, introduce the repeated measures correlation by conditioning the response on latent random effects, that is a finite mixture of random effects models. This is useful for instance to estimate time-varying unobserved heterogeneity as a mixture of discrete distribu-

tions or stochastic processes (Vermunt et al. 1999, Vermunt & Van Dijk 2001, Bartolucci & Farcomeni 2009, Bartolucci et al. 2014). These models rely on the local independence assumption, that is conditional on the cluster membership, the random effects and potential observed covariates, the subjects are assumed to be independent.

Vermunt & Van Dijk (2001) formulated a latent class regression model with class-specific coefficients, that is a finite mixture of random-intercepts and random-coefficients model. Given the lack of assumptions about the distribution of the random effects, the authors also viewed this model as a non-parametric two-level model. They applied this latent class regression model to responses with densities in the exponential family, including ordinal responses (Vermunt & Hagenaars 2004).

More recently, Bartolucci et al. (2014) presented a mixture of latent auto-regressive models for longitudinal binary, categorical and ordinal data that includes covariates as well as time-varying unobserved heterogeneity as a mixture of AR(1), autoregressive processes of order one, with different correlation coefficients but sharing the same variance. Frequentist estimation is performed using EM and Newtown-Raphson (NR) algorithms with sequential Gaussian quadrature to integrate out the mixture distribution of random effects. Model comparison is carried out using BIC and the S-index that takes into account the level of separation of the mixture components. They provide an application to self-reported health status for 7074 individuals over 8 years in the USA and taking into account the entropy based S-index argued for a model with three components, although a model with four components had the lowest BIC.

Latent Markov (LM) models (Wiggins 1973, Vermunt et al. 1999, Bartolucci et al. 2012) are another class of models that could also be considered parameter dependent. They are a generalization of finite mixtures where the cluster memberships arises from a discrete-state Markov chain and thus varies over time among the states. LM models are also known as Hidden Markov (HM) and Markov switching models in the time se-

ries literature (MacDonald & Zucchini 1997, Cappé et al. 2005, Zucchini & MacDonald 2009).

LM and HM models are more flexible than finite mixtures but also less parsimonious since additional parameters for the initial states and a transition matrix between states need to be estimated. Among the important contributions on this area, we highlight Bartolucci (2006) that provided a restricted likelihood ratio test (RLRT) to compare different configurations of the transition matrix for a given number of latent states, including whether or not the transition matrix is diagonal. They thus provided a test for the comparison of the LM and finite mixture formulations of models with the same number of latent states.

Ordinal data has been analysed using LM models by Vermunt et al. (1999) and Bartolucci & Farcomeni (2009). Vermunt et al. (1999) incorporates categorical time-constant and time-varying covariates to the LM for a categorical response, although they used the model for ordinal data. Importantly, in their formulation predictors affect initial and transition probabilities of the latent variable and not directly the response. The latter is also the case for Bartolucci & Farcomeni (2009), who proposed a multivariate extension of the dynamic logit model for binary, categorical and ordinal data. These authors used marginal logits for each response and marginal log-odds for each pair of responses and also allowed for covariates, including lagged responses. Their proposal moreover allows for time varying unobserved heterogeneity which is modelled as a first-order homogeneous Markov chain with a discrete number of states. The model is estimated using the EM and backward-forward recursions from the HM literature (MacDonald & Zucchini 1997) and model comparison is carried out using AIC and BIC. A simulation study shows how these information criteria worked well for the proposed model with BIC providing better results. They presented an application to fertility and employment as binary variables for 1446 women over 7 years. Using the AIC and the RLRT , they selected a LM model with three states which was preferred over the

corresponding finite mixture model. Interestingly, using the BIC rendered a different result as the model with the lowest overall BIC was a finite mixture mixture (and not a LM) with four components.

Chapter 5 presents a finite mixture model with latent random effects that encompasses the models of Vermunt & Van Dijk (2001) and Bartolucci et al. (2014) within one general framework. In particular, we propose a model with latent random effects that follow a standard normal distribution and Gaussian random walk process both cluster-specific variance. Our choice of sticking with finite mixtures, and not LM or HM models, was guided mainly by parsimony but also the aim of having both ways of modelling the time-varying unobserved heterogeneity within the same encompassing model. In contrast to these authors, we fit the models using a Bayesian approach and note their benefits and limitations in contrast to the Frequentist solutions.

1.5.2 Data dependent models

Data dependent models, introduce the repeated measures correlation by conditioning on a finite mixture of previous responses, that is by allowing the lagged response to have a different effect on each cluster. These models are also known as Markov transition or latent transition models and typically use time-homogeneous first-order Markov chains with states corresponding to the levels of the response(s). The latter is the key defining characteristic of this approach, which contrast with the LM/HM models where the Markov chain is defined over unobserved states. Markov transition models have been used for model based clustering of longitudinal data and time series (Frydman 2005, Pamminger et al. 2010, Frühwirth-Schnatter et al. 2012, Cheon et al. 2014).

In addition to the local independence assumption, models within this approach have to deal with the initial conditions problem (Heckman 1981*b*, Wooldridge 2005) due to their use of lagged responses as predictors. That

is, a joint model for the cluster membership and the response that occurred previous to the initial one also has to be specified. Recently, Skrondal & Rabe-Hesketh (2014) provide advice on the main approaches to tackle this issue.

Pamminger et al. (2010) and Frühwirth-Schnatter et al. (2012) present a mixture-of-experts Markov chain clustering, a model based clustering approach for categorical time series that uses a finite mixture of transitional terms and include covariates in the group membership probabilities. Known as "mixture-of-experts" in the machine learning literature, this model allows the covariate effects to be cluster specific and to deal with the initial conditions problem by adding the initial response into the set of regressors. This conditioning of the cluster membership on the initial response is an approach known as *simple solution* to the initial conditions problem in Econometrics (Wooldridge 2005). Model estimation is performed using MCMC and model selection with several Frequentist and Bayesian information criteria that take into account the entropy of the classification estimates. The models are illustrated using wage and income mobility in Austria. In both case studies, they found evidence for four latent groups with markedly different transitions over time. On the other hand, Frydman (2005), Cheon et al. (2014) developed restricted versions of the Markov transition models. Cheon et al. (2014) presents a disease progression model where the number of mixture components is equal to the disease states and thus is fixed in advance. Frydman (2005) considers another constrained model where the transition matrices for the latent groups are functions of the first group.

Similarly to Pamminger et al. (2010) and Frühwirth-Schnatter et al. (2012), the model proposed in Chapter 6 is a latent transition model that induces transition matrices that are completely unconstrained for all clusters. In contrast to them, the model we propose includes cluster-occasion interactions and thus the resulting transition matrices are time-varying. Furthermore, we construct the MCMC chain to sample from the target

distribution in a different manner, apply a different relabelling algorithm (Stephens 2000), and use the newly developed Widely Applicable Information Criterion (WAIC) (Watanabe 2009, 2010) for model comparison.

1.6 Thesis Roadmap

In the chapters to follow, this dissertation develops clustering models based on finite mixtures to attempt to fill some of these gaps. These probability models are based on likelihoods and thus provide a fuzzy clustering approach in which observations could come from any latent cluster with some probability. In particular, we have developed several finite mixture models for two-way data in longitudinal settings. To ease comparability, we start off by formulating models where the occasions are assumed to be independent, using mixtures based on the Proportional Odds and Trend Odds models in Chapters 3 and 4, respectively and fitting them using the EM algorithm. We then proceed to model the correlation explicitly with mixtures that include latent random effects in Chapter 5, and latent transitional terms in Chapter 6. These latter models are fitted using a Bayesian approach to take advantage of the flexibility of MCMC methods to estimate models with complex correlation structures. Furthermore, in Chapter 7 we use a Dirichlet Process prior to estimate the number of mixture components within a Bayesian Non-Parametric approach.

Throughout the dissertation, we validate the models using simulated data and also real data from socio-economic surveys and metagenomics. In particular, we used self-reported health status in Australia (poor to excellent), life satisfaction in New Zealand (completely agree to completely disagree) and site agreement with a reference genome (equal, segregating and variant) of bacteria from baby stool samples. More details of the these datasets are given next in Chapter 2.

Chapter 2

Data

2.1 Life Satisfaction (NZAVS)

The New Zealand Attitudes and Values survey (NZAVS) is a longitudinal survey hosted by the School of Psychology of the University of Auckland. It aims to study social attitudes, personality and health outcomes of New Zealanders. It was started in 2009 led by Associate Prof. Chris Sibley and now includes many researchers from a diverse range of research areas. Results and publication of all NZAVS data are independent of any specific funding agency or government body.

About to start its 7th year, the NZAVS is a postal survey planned to be a 20-year long study extending to 2029. The sample frame from this survey is drawn from the New Zealand Electoral Roll and started with 6,518 people in 2009. Since then it has had a average retention rate of around 80%. Including booster samples from 2011, its sample frame includes about 22,000 unique people. More technical information about the NZAVS could be found at www.psych.auckland.ac.nz/uoa/NZAVS

In this thesis, we use waves 1 to 5, 2009-2013, of self-reported "Life Satisfaction" (LS). Specifically, participants were asked the following:

The statements below reflect different opinions and points of view. Please indicate how strongly you disagree or agree with each statement. Remember, the best answer is your own opinion.

	Strongly Disagree						*Strongly Agree*
I am satisfied with my life	*1*	*2*	*3*	*4*	*5*	*6*	*7*

LS is therefore an ordinal variable with seven levels, ranging from 1 (Strongly disagree) to 7 (Strongly Agree). Given that we use all individuals with complete responses between 2009 and 2013, this dataset has 2564 rows (n), 5 columns (p) and 7 ordinal levels (q).

Over this period, most respondents exhibited very high levels of LS. As it can be seen in Figure 2.1, the majority of answers were very close to the highest end of the scale (5 to 7). In addition to that, there seems to be little variation of LS over time. Category 6 for instance was consistently just above 40% within this period. Similarly, categories 5 and 7 were around 20%. The remainder of categories all exhibited very low percentages, lower than 10% in the case of category 4 and lower than 5% in the case of categories 1,2 and 3.

More in detail, Table 2.1 shows the distribution of LS in 2009 and 2013 as well as the transitions between ordinal categories in this period. This table shows for instance that 54% of people that responded 7 (”Strongly Agree”) in 2009 also have the same response in 2013. In general, people with positive perception of their LS (”Agree” and ”Strongly Agree”) have diagonals that are higher than 50% which means that they tend to similar perceptions at the beginning and end of the study period.

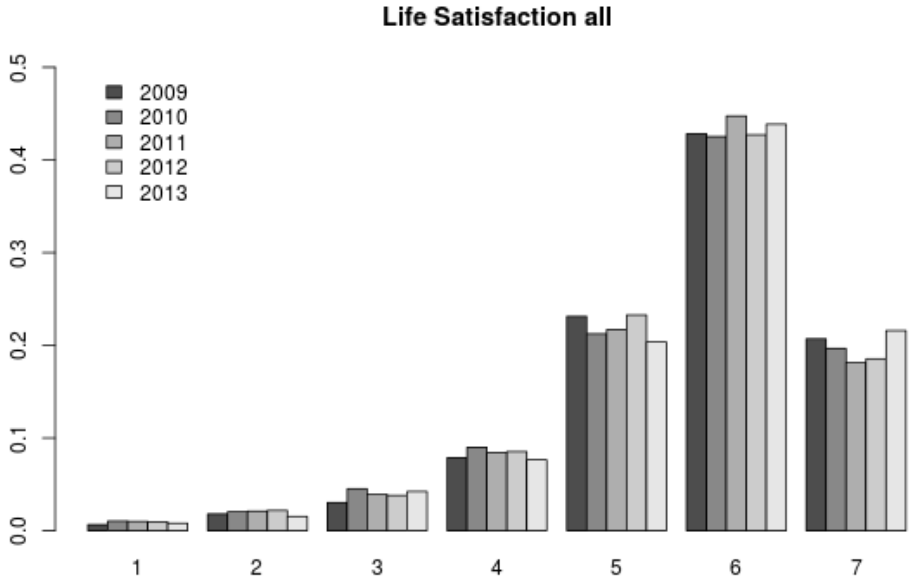


Figure 2.1: Distribution of Life Satisfaction (LS) over 2009-2013 in the NZAVS

Table 2.1: 2009-2013 transitions (%): Life Satisfaction in NZ

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
	2013	Strongly Disagree	Somewhat Disagree	Disagree	Neither	Somewhat Agree	Agree	Strongly Agree
2009		0.01	0.02	0.04	0.08	0.22	0.43	0.20
(1) Strongly Disagree	0.01	0.19	0.25	0.03	0.25	0.06	0.03	0.19
(2) Disagree	0.02	0.11	0.07	0.27	0.15	0.18	0.17	0.04
(3) Somewhat Disagree	0.04	0.02	0.17	0.20	0.24	0.21	0.15	0.02
(4) Neither	0.10	0.02	0.04	0.16	0.25	0.30	0.18	0.05
(5) Somewhat Agree	0.23	0.01	0.01	0.04	0.12	0.37	0.39	0.06
(6) Agree	0.39	0.00	0.00	0.01	0.03	0.17	0.60	0.18
(7) Strongly Agree	0.20	0.00	0.00	0.01	0.03	0.06	0.36	0.54

2.2 Health Satisfaction (HILDA)

The Household, Income and Labour Dynamics in Australia (HILDA) survey is a household panel study which began in 2001 and collects information about economic and subjective well-being, labour market dynamics and family dynamics in Australia. The HILDA Project was initiated and is funded by the Australian Government Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). Wave 1 in the panel had 7,682 households and 19,914 individuals and was topped up with an additional 2,153 households and 5,477 individuals in wave 11. More information about this survey can be found at: www.melbourneinstitute.com/hilda/

We use self-reported health status (SRHS) in 2001-2011 from this dataset. Using a five-level scale (Poor, Fair, Good, Very Good and Excellent) each year respondents answer the following question: "In general, would you say your health is:". We use individuals with complete records over 2001 to 2011, or 11 occasions. Therefore, for the HILDA dataset we have $n = 4660$ rows , $p = 11$ columns and $q = 5$ ordinal levels.

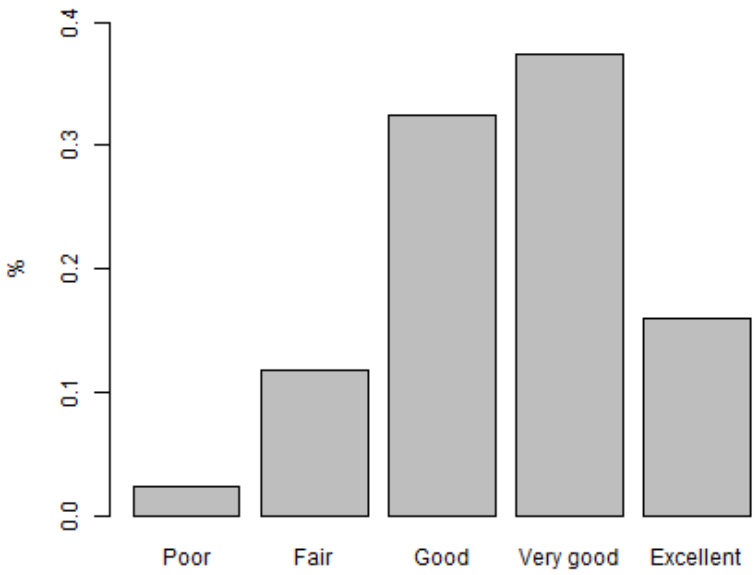
Figure 2.2 shows the distribution of SRHS in 2001 and 2011. In 2001, most individuals reported "Very Good" and "Good" health. About an eighth reported their health as "Excellent" and about a tenth as "Fair". A very low number of individuals said their health was "Poor". In contrast to that, in 2011 the same individuals reported lower health levels. "Excellent" and "Very Good" answers decreased and "Poor" and "Fair" increased. Overall, SRHS's distribution slightly shifted to the left and is thus more symmetric in 2011 than in 2001.

Furthermore, for each individual SRHS is highly correlated across time. Table 2.2 presents the 2001-2011 transitions between ordinal categories for all individuals. Diagonal proportions are very high, about 40%, and the same is true for the cells close to the diagonal. In words, even after 11 years individuals are very likely to report a very similar health status. The

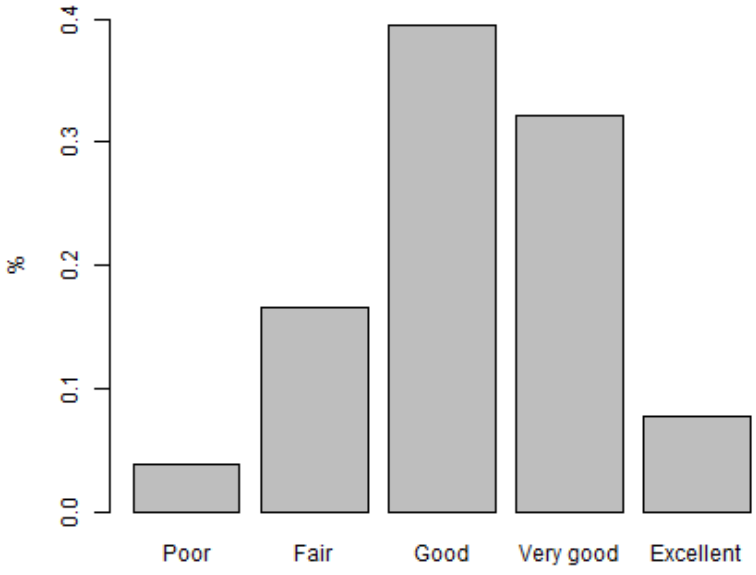
only exception to this, is people that responded "Excellent" in 2001. They have a slighter less positive perception of their health as 47% moved to "Very Good" over this period.

Table 2.2: SRHS transition matrix 2001-2011

		2011					
		Poor	Fair	Good	Very Good	Excellent	Total
2001	Poor	0.42	0.40	0.14	0.04	0.00	1.00
	Fair	0.13	0.44	0.34	0.07	0.01	1.00
	Good	0.02	0.21	0.54	0.20	0.02	1.00
	Very Good	0.01	0.09	0.38	0.46	0.07	1.00
	Excellent	0.01	0.04	0.21	0.47	0.27	1.00



2001



2011

Figure 2.2: Distribution of Self-Reported Health Status (SRHS) in 2001 and 2011 in HILDA

2.3 Infant gut bacteria (Metagenomics)

Metagenomics uses samples found in the physical environment to study genetic material. This contrasts with many other areas of Genomics where cultured samples are used. This dataset is a timeseries of infant gut bacterial composition that allows us to observe the developing of the infant gut. The main inferential goal when using this kind of data is to shed light on the dynamics of the (latent) strains of bacteria (b.faecis) that compete to inhabit the human gut after birth. Figure 2.3 shows the sampling timeline.

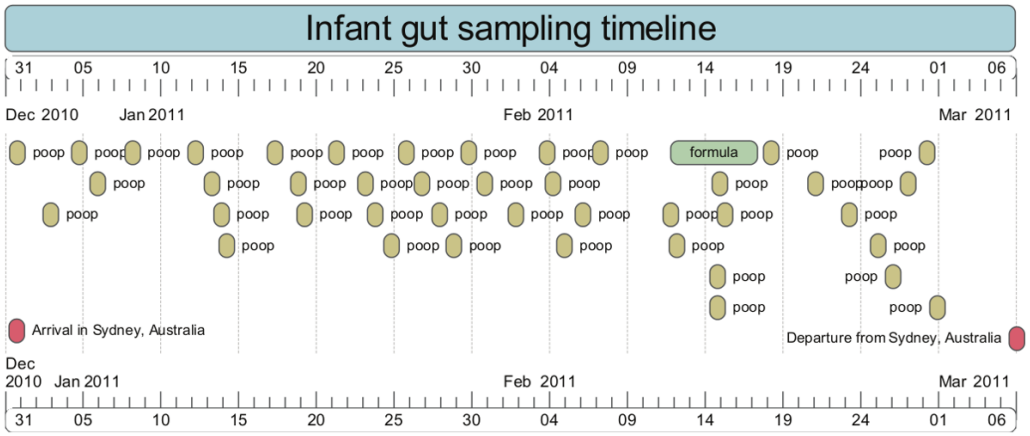


Figure 2.3: Infat gut sampling timeline. Source: Chan et al. (2015)

More in detail, this dataset consists of reference and variant allele counts for 62,996 single-nucleotide variant (SNV) sites, specific positions in the genome of the bacteria, followed over 45 days. These counts are then used to produce and ordinal variable, infant gut bacteria variants, with three levels:"fixed to reference", "segregating site", and "fixed to a non-reference". At each time point, these levels are defined as follows

- "fixed to reference": includes SNV sites where all reads are the same as the reference;
- "segregating site" more than 5 reads are equal to reference and more

than 5 reads are equal to a different one;

- "fixed to non-reference" all reads for the cell are the same allele which is different to the reference one.

Figure 2.4 shows a heatmap of the data using yellow, red and blue for these levels. Fitting a Poisson mixture for the reads with Automatic Differentiation Variational Inference, Chan et al. (2015) found evidence for at least three strains, that is at least three different patterns of reads overtime.

Given the large size of this dataset and the amount of missing observations, sites with no reads at all, we use all observations with complete data over the first 25 occasions. Thus, the infant gut dataset has $n = 1992$ rows, $p = 25$ columns and $q = 3$ ordinal levels and is publicly available at my personal repository: bitbucket.org/cholokiwi/nzsa2016/src.

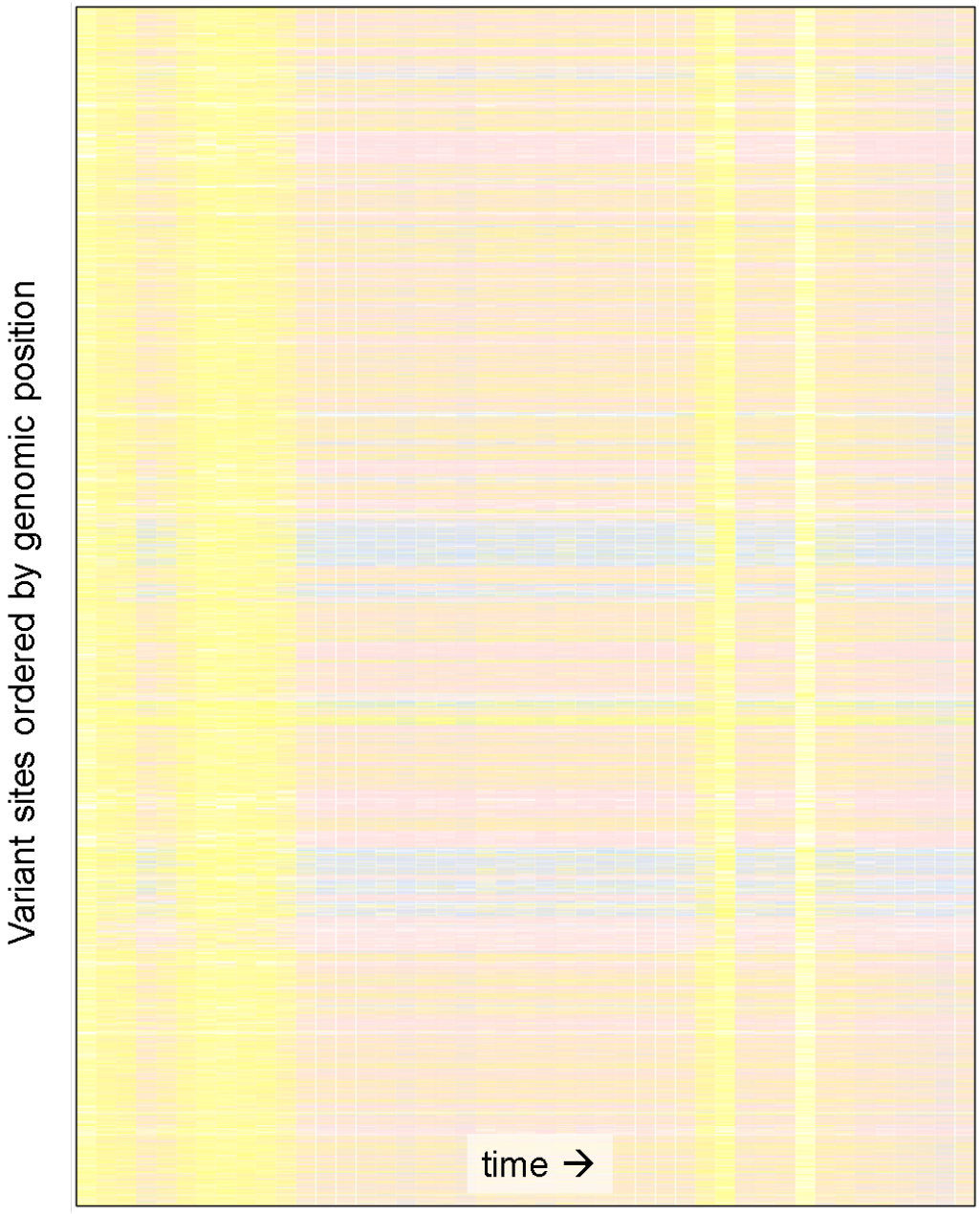


Figure 2.4: Heatmap for infant gut bacteria variants over 45 days. Source: Chan et al. (2015)

Part I

Frequentist Estimation

Chapter 3

Proportional Odds Model

3.1 The Model

In this chapter, we extend the Proportional Odds Model (POM) by McCullagh (1980) to the case of latent groups. That is, we introduce unobserved covariates into the linear predictor of the cumulative probability of observing the ordinal outcomes. We already introduced this model in the previous chapter (Section 1.2). As a starting point, the models in this chapter assume that the observations are independent over time. By doing so, we thus use the same framework as Matechou et al. (2016).

The setup that follows will be used throughout the present as well as in Chapters 3, 4, 5 and 6. Let data Y be a (n, p) matrix where each cell y_{ij} is equal to any of the q ordinal categories, where: $i = 1, ..., n$; $j = 1, ..., p$ and $k = 1, ..., q$. This is, each response y_{ij} is the realization of a multinomial distribution with probabilities $\theta_{ij1}, \ldots \theta_{ijq}$, where $\theta_{ijk} \geq 0$ and $\sum_{k=1}^{q} \theta_{ijk} = 1$. We also define an indicator variable $I(y_{ij} = k)$ equal to 1 if the condition $y_{ij} = k$ is satisfied and 0 otherwise. v represents the number of model parameters.

Saturated model

We begin by formulating a model where every single row (i) and column (j) and its interactions have an effect in the linear predictor of the POM, the saturated model

$$\text{Logit}[P(y_{ij} \leq k)] = \mu_k - \alpha_i - \beta_j - \gamma_{ij} \tag{3.1}$$

where $i = 1 \dots n$, $j = 1 \dots p$, $k = 1, \dots, q$ and the following identifiability constraints:

$$\begin{aligned}
&\alpha_1 = 0 \\
&\beta_1 = 0 \\
&\gamma_{1j} = 0, \forall j; \gamma_{i1} = 0, \forall i \\
&\mu_{k-1} < \mu_k, \ k = 1, \dots, (q-1) \text{ and } \mu_0 = -\infty, \mu_q = \infty
\end{aligned}$$

The parameter μ_k is the k^{th} cut point, α_i is the effect of row i, β_j is the effect of column j and γ_{ij} is an interaction effect between row i and column j. The model can also be expressed in terms of the probabilities of each ordinal outcome θ_{ijk}

$$P(y_{ij} = k) = \theta_{ijk} = \frac{1}{1 + e^{-(\mu_k - \alpha_i - \beta_j - \gamma_{ij})}} - \frac{1}{1 + e^{-(\mu_{k-1} - \alpha_i - \beta_j - \gamma_{ij})}}$$

This model is not at all parsimonious as it has $v = (q-1) + (n-1) + (p-1) + (n-1)(p-1)$ parameters.

Row and column effects model

A more parsimonious alternative is the model with only main row (i) and column (j) effects

$$\text{Logit}[P(y_{ij} \leq k)] = \mu_k - \alpha_i - \beta_j \tag{3.2}$$

or

$$P(y_{ij} = k) = \theta_{ijk} = \frac{1}{1 + e^{-(\mu_k - \alpha_i - \beta_j)}} - \frac{1}{1 + e^{-(\mu_{k-1} - \alpha_i - \beta_j)}}$$

This model has the same constraints on μ, α, β as above and it has $v = (q-1)+(n-1)+(p-1)$ parameters. v is still potentially large and increases linearly with the sample size. More parsimonious alternatives are thus needed.

Row-clustering only

Suppose now that each row belongs to one of the $r = 1, \ldots, R$ row groups with probabilities $\pi_1, \ldots, \pi_R$. That is, we assume that the rows come from a finite mixture with R components where both R and the group membership r_i are unknown. Note that $R < n$ and $\pi_r \geq 0, \sum_{r=1}^{R} \pi_r = 1, \forall i$.

Now, let θ_{rjk} be the probability that observation y_{ij} equals ordinal category k given that row i belongs to row-cluster r: $P(y_{ij} = k | i \in r) = \theta_{rjk}$. In this simple case, row-clustering only with no column effects, the model is:

$$\text{Logit}[P(y_{ij} \leq k | i \in r)] = \mu_k - \alpha_r \tag{3.3}$$

which implies

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

Here $\alpha_1 = 0$, and $\mu_{k-1} < \mu_k,\ k = 1 \ldots (q-1)$, $\mu_0 = -\infty$ and $\mu_q = \infty$. μ_k is the k^{th} cut-off point and α_r is the effect of row-cluster r.

Assuming independence over the rows and, conditional on the rows, independence over the columns, the likelihood becomes

$$L(\phi, \pi | Y) = \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)} \tag{3.4}$$

where $\phi = (\mu, \alpha)$ is the set of parameters in the linear predictor. The expression above is also referred as the incomplete data likelihood given that

the cluster memberships of the rows are unknown. The number of model parameters is equal to: $v = (q-1)+2(R-1)$. Note that the linear predictor in this simple case does not depend on the columns (j) and thus we could have also written θ_{rjk} as θ_{rk}.

Row-clustering with column effects

To introduce column effects we augment the linear predictor to

$$\text{Logit}[P(y_{ij} \leq k|i \in r)] = \mu_k - \alpha_r - \beta_j \tag{3.5}$$

which implies

$$P(y_{ij} = k|i \in r) = \theta_{rjk} = \frac{1}{1+e^{-(\mu_k-\alpha_r-\beta_j)}} - \frac{1}{1+e^{-(\mu_{k-1}-\alpha_r-\beta_j)}} \tag{3.6}$$

where $\alpha_1 = \beta_1 = 0$, and $\mu_{k-1} < \mu_k,\ k = 1 \ldots (q-1)$. μ_k is the k^{th} cut-off point, α_r is the effect of row-cluster r and β_j the effect of column j. In this case, the likelihood is

$$L(\phi, \pi|Y) = \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)} \tag{3.7}$$

where: ϕ=(μ, α, β) and the number of parameters in the model is $v = (q-1) + 2(R-1) + (p-1)$.

Row-clustering with column effects and interactions

To introduce column effects and interactions in the row-clustered model, we further augment the linear predictor to

$$\text{Logit}[P(y_{ij} \leq k|i \in r)] = \mu_k - \alpha_r - \beta_j - \gamma_{rj} \tag{3.8}$$

which is equivalent to

$$P(y_{ij} = k|i \in r) = \theta_{rjk} = \frac{1}{1+e^{-(\mu_k-\alpha_r-\beta_j-\gamma_{rj})}} - \frac{1}{1+e^{-(\mu_{k-1}-\alpha_r-\beta_j-\gamma_{rj})}}$$

Where $\alpha_1 = \beta_1 = 0$; $\gamma_{1j} = 0, \forall j$; $\gamma_{r1} = 0, \forall r$; and $\mu_k > \mu_{k-1},\ k = 1 \dots (q-1)$. μ_k is the k^{th} cut-off point, α_r is the effect of row-cluster r, β_j the effect of column j and γ_{rj} the row-cluster and column interaction. The likelihood for this model is

$$L(\phi, \pi|Y) = \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)} \tag{3.9}$$

where: $\phi = (\mu, \alpha, \beta, \gamma)$ and the number of model parameters $v = (q-1) + 2(R-1) + (p-1) + (R-1)(p-1)$.

Column-clustering only

Column-clustering is also one-way clustering and is equivalent to row-clustering with the transposed data. That is, we could obtain column groups by exchanging row and columns and applying the row-cluster model from the previous section. Setting C as the number of mixture components, κ_c as the mixture proportion for group c, and $P(y_{ij} = k|j \in c) = \theta_{ick}$, the model becomes

$$\text{Logit}[P(y_{ij} \leq k|j \in c)] = \mu_k - \beta_c \tag{3.10}$$

or alternatively

$$P(y_{ij} = k|j \in c) = \theta_{ick} = \frac{1}{1 + e^{-(\mu_k - \beta_c)}} - \frac{1}{1 + e^{-(\mu_{k-1} - \beta_c)}}$$

where $\beta_1 = 0$, and $\mu_{k-1} < \mu_k,\ k = 1 \dots (q-1)$. μ_k is the k^{th} cut-off point and β_c is the effect of column-cluster c. Note also that $C < p$ and $\kappa_c \geq 0, \sum_{c=1}^{C} \kappa_c = 1$. Assuming independence over the columns and independence over the rows conditional on the columns, the model's likelihood is

$$L(\phi, \kappa|Y) = \prod_{j=1}^{p} \sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{ick}^{I(y_{ij}=k)} \tag{3.11}$$

where $\phi = (\mu, \beta)$. The incomplete information likelihood for this only column-cluster case has $v = (q-1) + 2(C-1)$ parameters.

Column-clustering with row-effects

We introduce row-effects to the column-clustering by augmenting the model in (3.10) to

$$\text{Logit}[P(y_{ij} \leq k|j \in c)] = \mu_k - \beta_c - \alpha_i \tag{3.12}$$

or alternatively

$$P(y_{ij} = k|j \in c) = \theta_{ick} = \frac{1}{1 + e^{-(\mu_k - \beta_c - \alpha_i)}} - \frac{1}{1 + e^{-(\mu_{k-1} - \beta_c - \alpha_i)}}$$

where $\beta_1 = \alpha_1 = 0$, and $\mu_{k-1} < \mu_k,\ k = 1 \ldots (q-1)$. μ_k is the k^{th} cut-off point, β_c is the effect of column-cluster c and α_i the effect of row i. Note also that $C < p$ and $\sum_{c=1}^{C} \kappa_c = 1$. Its likelihood is then

$$L(\phi, \kappa|Y) = \prod_{j=1}^{p} \sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{ick}^{I(y_{ij}=k)} \tag{3.13}$$

where $\phi = (\mu, \beta, \alpha)$ and $v = (q-1) + 2(C-1) + (n-1)$ parameters.

Column-clustering with row-effects and interactions

In this case, the model is described by

$$\text{Logit}[P(y_{ij} \leq k|j \in c)] = \mu_k - \beta_c - \alpha_i - \gamma_{ic} \tag{3.14}$$

or alternatively

$$P(y_{ij} = k|j \in c) = \theta_{ick} = \frac{1}{1 + e^{-(\mu_k - \beta_c - \alpha_i - \gamma_{ic})}} - \frac{1}{1 + e^{-(\mu_{k-1} - \beta_c - \alpha_i - \gamma_{ic})}}$$

here $\beta_1 = \alpha_1 = 0$; $\gamma_{i1} = 0, \forall i$; $\gamma_{1c} = 0, \forall c$; and $\mu_{k-1} < \mu_k,\ k = 1 \ldots (q-1)$. μ_k is the k^{th} cut-off point, β_c is the effect of column-cluster c, α_i the effect of row i and γ_{ic} the column-cluster and row interaction. The likelihood for this model is

$$L(\phi, \kappa|Y) = \prod_{j=1}^{p} \sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{ick}^{I(y_{ij}=k)} \tag{3.15}$$

with $\phi = (\mu, \beta, \alpha, \gamma)$ and $v = (q-1) + 2(C-1) + (n-1) + (C-1)(n-1)$ parameters.

Bi-clustering

Bi-clustering is also known as two-way, two-mode and latent block clustering (Govaert & Nadif 2010, Pledger & Arnold 2014, Matechou et al. 2016) and it consists of simultaneous clustering of the rows and columns. Here, it is assumed that rows come from a finite mixture with R row groups while the columns come from a finite mixture with C column groups, simultaneously. The row and column-cluster proportions are $\pi_1, \ldots, \pi_R$ and $\kappa_1, \ldots, \kappa_C$, respectively. R, C, π_r and κ_c are unknown. Note that $R < n$, $C < p$, $\sum_{r=1}^{R} \pi_r = 1$ and $\sum_{c=1}^{C} \kappa_c = 1$.

Let $\theta_{rck} = P(y_{ij} = k | i \in r \cap j \in c)$ be probability that cell (i, j) is equal to ordinal outcome k given that it belongs to row group r and column group c. Keeping as before α_r and β_c, for the row and column-cluster effects, the linear predictor becomes:

$$\text{Logit}[P(y_{ij} \leq k | i \in r \cap j \in c)] = \mu_k - \alpha_r - \beta_c \tag{3.16}$$

or alternatively

$$P(y_{ij} = k | i \in r \cap j \in c) = \theta_{rck} = \frac{1}{1 + e^{-(\mu_k - \alpha_r - \beta_c)}} - \frac{1}{1 + e^{-(\mu_{k-1} - \alpha_r - \beta_c)}}$$

Where $\alpha_1 = \beta_1 = 0$ and $\mu_{k-1} > \mu_k,\ k = 1 \ldots (q-1)$, $\mu_0 = -\infty$ and $\mu_q = \infty$.

In this case, the likelihood sums over all possible partitions of rows into R clusters and over all possible partitions of columns into C clusters. Assuming independence over the rows and independence over the columns conditional on the rows, the incomplete data likelihood could be simplified to

$$L(\phi, \pi, \kappa | Y) = \sum_{c_1=1}^{C} \cdots \sum_{c_p=1}^{C} \kappa_{c_1} \ldots \kappa_{c_p} \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rc_jk}^{I(y_{ij}=k)} \tag{3.17}$$

where $\phi = (\mu, \alpha, \beta)$. This expression is computationally expensive to evaluate since it requires consideration of all possible allocations of the p columns to the C groups. Alternatively, assuming independence over the columns and independence over the rows conditional on the columns, it simplifies to

$$L(\phi, \pi, \kappa|Y) = \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \ldots \pi_{r_n} \prod_{j=1}^{p} \sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{r_i ck}^{I(y_{ij}=k)} \tag{3.18}$$

Likewise (3.17), (3.18) is very expensive to compute due to requiring consideration of all possible allocations of the n rows to the R groups. In either specification, the incomplete data likelihood for the bi-clustering has $v = (q-1) + 3(R + C - 2)$ parameters.

Bi-clustering with interactions

In this case, the model is

$$\text{Logit}[P(y_{ij} \leq k|i \in r \cap j \in c)] = \mu_k - \alpha_r - \beta_c - \gamma_{rc} \tag{3.19}$$

or alternatively

$$P(y_{ij} = k|i \in r \cap j \in c) = \theta_{rck} = \frac{1}{1 + e^{-(\mu_k - \alpha_r - \beta_c - \gamma_{rc})}} - \frac{1}{1 + e^{-(\mu_{k-1} - \alpha_r - \beta_c - \gamma_{rc})}}$$

Where $\alpha_1 = \beta_1 = 0$; $\gamma_{1c} = 0, \forall c$; $\gamma_{r1} = 0, \forall r$; and $\mu_k > \mu_{k-1},\ k = 1 \ldots (q-1)$, $\mu_0 = -\infty$ and $\mu_q = \infty$. μ_k is the k^{th} cut-off point, α_r is the r row-cluster effect, β_c is the c column-cluster effect and γ_{rc} the row-cluster and column-cluster interaction.

Depending on the independence assumptions, over the rows conditional on the columns or vice versa, the likelihood for this model is the same as in (3.17) or (3.18) with the augmented linear predictor $\phi = (\mu, \alpha, \beta, \gamma)$.

3.2 Frequentist Estimation

In this section we use a Frequentist framework to maximise the likelihoods described above and estimate the model parameters. In particular we use the Expectation-Maximization (EM) algorithm (Dempster et al. 1977). In essence, the EM algorithm turns the estimation into a missing data problem and then estimates the parameters using an iterative two-fold approach. In the case of finite mixtures, it first introduces new variables for the unknown group memberships. It then initialises these missing data given some initial values for the parameters (E-step). Secondly, the parameters are estimated by maximizing the likelihood given the estimated missing data (M-step). This likelihood is known as "complete data" likelihood since is assumes that the latent group memberships are known. The new parameters in turn are fed to the E-step again and the process repeats until the parameters converge, that is when the change in the parameters and/or the likelihood is tiny.

Row-clustering

Let z_{ir} be the latent row group memberships for each row. z_{ir} is an indicator function equal to 1 if row i belongs to cluster r and 0 otherwise. It is not observed and thus is regarded as missing data. Note that $\sum_{r=1}^{R} z_{ir} = 1, \forall i$, and we define a membership matrix $Z_{(n,R)}$ to gather together all the z_{ir}'s. As before ϕ denotes the linear predictor parameters so that the parameters of the model are (ϕ, π). Given a value for the number of mixture components R, the EM algorithm proceeds as follows.

E-step: Estimate Z. Given Y and initial values for ϕ and π, estimate the expected value of z_{ir} as

$$E[z_{ir}|Y, \phi, \pi] = \hat{z}_{ir} = \frac{\pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)}}{\sum_{a=1}^{R} \pi_a \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{ajk}^{I(y_{ij}=k)}} \tag{3.20}$$

In words, $\hat{z}_{ir}$ is the posterior probability that observation y_{ij} belongs to the (row) mixture component r. $\sum_{r=1}^{R} \hat{z}_{ir} = 1, \forall i$.

M-step: Numerically maximise the complete data log-likelihood. Given $\hat{z}_{ir}$ from the E-step maximise ℓ_c to obtain new values for ϕ and π

$$\ell_c(\phi, \pi|Y, Z) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{k=1}^{q} \hat{z}_{ir} I(y_{ij} = k) \log(\theta_{rjk}) + \sum_{i=1}^{n}\sum_{r=1}^{R} \hat{z}_{ir} \log(\hat{\pi}_r) \tag{3.21}$$

where $\hat{\pi}_r = E[\pi_r|Z] = \sum_{i=1}^{n} \hat{z}_{ir}/n$. That is, the mixing proportions $\hat{\pi}_r$ are first obtained using Z.

A new cycle starts when the parameters from the M-step are used in the E-step and the Z matrix is re-estimated. This process repeats until estimates for ϕ have converged.

The parameters in the linear predictor depend on type of clustering we are performing. For the row-clustering case we have $\phi = (\mu, \alpha)$, for row-clustering with column effects $\phi = (\mu, \alpha, \beta)$ and for row-clustering with column effects and interactions $\phi = (\mu, \alpha, \beta, \gamma)$.

Importantly, as with any other maximization algorithm, there is a risk of convergence to local maxima due to multimodality of the likelihood. It is thus important to start the EM algorithm with several widely spread initial values, McLachlan & Peel (2000), and check that all of them converge to the same place. (log-likelihood value)

Column-clustering

Let x_{jc} be the latent column-cluster memberships for each column j and X the membership matrix where each cell is equal to x_{jc}, $\sum_{c=1}^{C} x_{jc} = 1$. Given a value for the number of mixture components C, the EM algorithm proceeds as follows.

E-step: Estimate X. Given Y and initial values for ϕ, we estimate the expected values of x_{jc} as

$$E[x_{jc}|Y,\phi,\kappa] = \hat{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n}\prod_{k=1}^{q}\theta_{ick}^{I(y_{ij}=k)}}{\sum_{a=1}^{C}\kappa_a \prod_{i=1}^{n}\prod_{k=1}^{q}\theta_{iak}^{I(y_{ij}=k)}} \tag{3.22}$$

That is, $\hat{x_{jc}}$ is the posterior probability that observation y_{ij} belongs to the (column) mixture component c.

M-step: Numerically maximise the complete data log-likelihood. Given $\hat{x}_{ir}$ from the E-step maximise ℓ_c to obtain new values for ϕ and κ

$$\ell_c(\phi,\kappa|Y,X) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{c=1}^{C}\sum_{k=1}^{q}\hat{x}_{jc}I(y_{ij}=k)\log(\theta_{ick}) + \sum_{j=1}^{p}\sum_{c=1}^{C}\hat{x}_{jc}\log(\hat{\kappa}_c) \tag{3.23}$$

where $\hat{\kappa}_c = E[\kappa_c|X] = \sum_{j=1}^{p}\hat{x}_{jc}/p$. That is, the mixing proportions $\hat{\kappa_c}$ are first obtained using X.

A new cycle starts when the parameters from the M-step are use in the E-step. This process repeats until convergence for ϕ. Again, the parameters in the linear predictor depend on type of clustering used: $\phi = (\mu,\alpha)$, $\phi = (\mu,\alpha,\beta)$, and $\phi = (\mu,\alpha,\beta,\gamma)$ for the column-clustering, column-clustering with row effects, and column-clustering with row effects and interactions, respectively.

Bi-clustering

Let z_{ir} and x_{jc} be the latent row and column cluster memberships for each cell (i,j). As before Z and X are membership matrices formed by all the z_{ir}, x_{jc} values. Note that $\sum_{r=1}^{R}x_{ir} = \sum_{c=1}^{C}x_{jc} = 1$. Let ϕ be the set of model parameters for the biclustering case, we also incorporate the variational approximation employed by Govaert & Nadif (2005)

$$E[z_{ir}x_{jc}|Y,\phi] \simeq E[z_{ir}|Y,\phi]E[x_{jc}|Y,\phi] = \hat{z}_{ir}\hat{x}_{jc}$$

That is, conditonal on the ordinal response and the parameters the effect of the row and column clusters are independent. Given this approxi-

mation, values for the number of mixture components (R, C), and assuming independence over the rows and independence over the columns conditional on the rows, the EM algorithm proceeds as follows.

E-step: Estimate Z and X. Given Y and initial values for ϕ estimate the expected values of z_{ir} and x_{jc} as

$$\begin{aligned} E[z_{ir}|Y,\phi,\pi] = \hat{z}_{ir} &= \frac{\pi_r \prod_{j=1}^{p} \left\{ \sum_{c=1}^{C} \kappa_c \prod_{k=1}^{q} \theta_{rck}^{I(y_{ij}=k)} \right\}}{\sum_{a=1}^{R} \pi_a \prod_{j=1}^{p} \left\{ \sum_{b=1}^{C} \kappa_b \prod_{k=1}^{q} \theta_{abk}^{I(y_{ij}=k)} \right\}} \\ E[x_{jc}|Y,\phi,\kappa] = \hat{x}_{jc} &= \frac{\kappa_c \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \pi_r \prod_{k=1}^{q} \theta_{rck}^{I(y_{ij}=k)} \right\}}{\sum_{b=1}^{C} \kappa_b \prod_{i=1}^{n} \left\{ \sum_{a=1}^{R} \pi_a \prod_{k=1}^{q} \theta_{abk}^{I(y_{ij}=k)} \right\}} \end{aligned} \tag{3.24}$$

M-step: Numerically maximise the complete data log-likelihood. Given $\hat{z}_{ir}$ and $\hat{x}_{jc}$ from the E-step maximise ℓ_c to obtain new values for ϕ, π, κ

$$\begin{aligned} \ell_c(\phi,\pi,\kappa|Y,Z,X) = &\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{q} \hat{z}_{ir}\hat{x}_{jc} I(y_{ij}=k)\log(\theta_{rck}) \\ &+ \sum_{i=1}^{n}\sum_{r=1}^{R} \hat{z}_{ir}\log(\hat{\pi}_r) \\ &+ \sum_{j=1}^{p}\sum_{c=1}^{C} \hat{x}_{jc}\log(\hat{\kappa}_c) \end{aligned} \tag{3.25}$$

where: $\hat{\pi_r} = E[\pi_r|Z] = \sum_{i=1}^{n} \hat{z}_{ir}/n$ and $\hat{\kappa_c} = E[\kappa_c|X] = \sum_{j=1}^{p} \hat{x}_{jc}/p$.

A new cycle starts when the parameters from the M-step are use in the E-step. This process repeats until convergence. Note that θ_{rck} is estimated using the corresponding linear predictor, bi-clustering in (3.16) and bi-clustering with interactions in (3.19).

3.3 Simulations

In this chapter, we use synthetic datasets to illustrate the performance of the models presented so far. In particular, we simulate several scenarios and evaluate the performance of the POM model with row-clustering and column effects (3.5). This model is more complex than the POM model with row-clustering but more parsimonious than the POM model with row-clustering with column effects and interactions.

We generate a total of 18 scenarios with varying sample size (3), number of columns (2) and mixture proportions (3). We use the following setup:

- $R = 3$ number of mixture components
- $\mu = (-2.08, -1.39, 1.39, 2.08)$ cut points
- $q = 5$ ordinal levels,
- $\alpha = (0, -2, 3)$ cluster effects
- $\beta = (0, 0.15, 0.30, \ldots, 0.15(p-1))$ column effects.

Given this setup, the model has $v = 12$ and 17 parameters, for $p = 5$ and 10, respectively. Table 3.1 details these scenarios

Table 3.1: Labels for the scenarios (1-18) in the simulation study

	$p = 5$			$p = 10$		
	n			n		
	60	200	1000	60	200	1000
$\pi = (0.50, 0.30, 0.20)$	1	2	3	4	5	6
$\pi = (0.33, 0.33, 0.34)$	7	8	9	10	11	12
$\pi = (0.90, 0.08, 0.02)$	13	14	15	16	17	18

For each scenario, we generate 500 simulated datasets with 20 random initial values each to avoid the risk of convergence to local maxima. We

therefore estimate $3 \times 2 \times 3 \times 500 \times 20 = 180,000$ models overall. We estimate the model parameters using the EM algorithm as detailed in Section 3.2. Convergence is set to a maximum change of 10^{-8} in the absolute value of all the parameter estimates between EM iterations. The models are implemented in R (R 3.2.2) and C++. In particular, estimation in the M-step is carried out using the *optim()* function in R and the corresponding log-likelihoods in C++ compiled functions. Tables 3.2 to 3.8 and figures 3.1 to 3.4 present the results.

We start with scenarios 1-6 in Tables 3.2 and 3.3 where the mixture proportions $\pi = (0.20, 0.30, 0.50)$ are unbalanced but the smallest component is not too small. Overall, we could see that the means of estimated parameters are close to their true values for all the model parameters. That is, the estimates for the cut-off points μ, cluster effects α, column effects β, and mixture proportions π all seem to converge to their true values. The standard deviations of these estimates are large for the $n = 60$ sample size. This is not surprising since we are estimating $v = 12$ independent parameters with only 60 rows and 5 columns. Increasing n has the effect of both providing estimates closer to their true values and decreasing the standard errors. For $n = 1000$ the mean values are already very close the true ones with small standard errors. Scenarios 4-6 present the estimates with twice the number of columns ($p = 10$ instead of $p = 5$). The model in this case has $v = 17$ independent parameters. Table 3.3 shows that the mean values of the estimates are close to the true parameter values with decreasing standard deviations in n. Furthermore, these estimates exhibit lower variability than the previous ones.

Looking closely into the estimates of cluster effects α_r, Figure 3.1 presents all the estimates for α over the 500 replicated datasets. It is evident that bigger n provides closer estimates to the true value for α. However, there seem to be some estimates that are very different from the true values. For example for $n = 60$ and $p = 5$ there are estimates near the axis and thus are clearly far off from the true values $(-2, 3)$ in the middle of the graph.

Importantly, these outliers disappear when n or p increase.

Table 3.2: Scenarios 1-3. Estimated parameters for the POM with row-clustering and column effects. Mean and standard deviation (sd) over 500 simulated datasets. $\pi = (0.50, 0.30, 0.20)$ and $p = 5$.

		$n = 60$		$n = 200$		$n = 1000$	
Param	true	mean	sd	mean	sd	mean	sd
μ_1	-2.08	-2.21	0.58	-2.10	0.24	-2.08	0.10
μ_2	-1.39	-1.49	0.58	-1.40	0.23	-1.39	0.10
μ_3	1.39	1.34	0.58	1.40	0.21	1.39	0.09
μ_4	2.08	2.05	0.60	2.11	0.22	2.09	0.10
α_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
α_2	-2.00	-2.16	0.40	-2.05	0.19	-2.00	0.09
α_3	3.00	3.04	0.65	3.05	0.24	3.01	0.11
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_2	0.15	0.18	0.38	0.15	0.20	0.15	0.09
β_3	0.30	0.30	0.39	0.30	0.22	0.30	0.09
β_4	0.45	0.48	0.37	0.46	0.20	0.45	0.09
β_5	0.60	0.64	0.39	0.61	0.20	0.60	0.09
π_1	0.50	0.48	0.12	0.50	0.06	0.50	0.03
π_2	0.30	0.32	0.13	0.30	0.06	0.30	0.03
π_3	0.20	0.20	0.06	0.20	0.03	0.20	0.01

Table 3.3: Scenarios 4-6. Estimated parameters for the POM with row-clustering and column effects. Mean and standard deviation (sd) over 500 simulated datasets. $\pi = (0.50, 0.30, 0.20)$ and $p = 10$.

		$n = 60$		$n = 200$		$n = 1000$	
Param	true	mean	sd	mean	sd	mean	sd
μ_1	-2.08	-2.12	0.33	-2.09	0.18	-2.08	0.07
μ_2	-1.39	-1.42	0.31	-1.39	0.17	-1.39	0.07
μ_3	1.39	1.40	0.29	1.40	0.16	1.39	0.07
μ_4	2.08	2.10	0.30	2.10	0.17	2.08	0.07
α_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
α_2	-2.00	-2.04	0.22	-2.01	0.11	-2.00	0.05
α_3	3.00	3.05	0.31	3.03	0.16	3.00	0.07
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_2	0.15	0.17	0.35	0.16	0.21	0.15	0.09
β_3	0.30	0.28	0.36	0.30	0.20	0.30	0.09
β_4	0.45	0.46	0.37	0.45	0.20	0.45	0.09
β_5	0.60	0.63	0.38	0.60	0.19	0.59	0.09
β_6	0.75	0.76	0.38	0.76	0.21	0.74	0.09
β_7	0.90	0.91	0.36	0.92	0.21	0.90	0.09
β_8	1.05	1.06	0.37	1.06	0.20	1.05	0.09
β_9	1.20	1.20	0.37	1.22	0.20	1.20	0.09
β_{10}	1.35	1.37	0.36	1.35	0.21	1.35	0.09
π_1	0.50	0.50	0.08	0.50	0.04	0.50	0.02
π_2	0.30	0.31	0.07	0.30	0.04	0.30	0.02
π_3	0.20	0.20	0.05	0.20	0.03	0.20	0.01

Tables 3.4 and 3.5 present the results for scenarios 7-12, which have balanced mixture proportions $\pi = (0.33, 0.33, 0.34)$. In general, the means of estimated parameters are also close to their true values for all the model parameters. However, the main difference of these new scenarios is that the variability of these estimates (sd) is very similar among all the mixture components. This is somewhat expected as all the mixture components are equally represented in the data. On the other hand, occasional outliers seem to be also present in the estimates of the cluster effects α, Figure 3.2, but they dissapear as n and p grow.

Table 3.4: Scenarios 7-9. Estimated parameters for the POM with row-clustering and column effects. Mean and standard deviation (sd) over 500 simulated datasets. $\pi = (0.33, 0.33, 0.34)$ and $p = 5$.

		$n = 60$		$n = 200$		$n = 1000$	
Param	true	mean	sd	mean	sd	mean	sd
μ_1	-2.08	-2.36	0.75	-2.11	0.30	-2.08	0.13
μ_2	-1.39	-1.65	0.76	-1.41	0.30	-1.39	0.13
μ_3	1.39	1.18	0.77	1.38	0.27	1.39	0.12
μ_4	2.08	1.91	0.77	2.07	0.28	2.08	0.12
α_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
α_2	-2.00	-2.31	0.77	-2.05	0.22	-2.00	0.09
α_3	3.00	3.01	0.74	3.02	0.24	3.00	0.10
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_2	0.15	0.21	0.36	0.15	0.21	0.15	0.09
β_3	0.30	0.33	0.36	0.30	0.22	0.30	0.09
β_4	0.45	0.48	0.37	0.45	0.21	0.45	0.08
β_5	0.60	0.66	0.38	0.62	0.21	0.60	0.09
π_1	0.33	0.32	0.11	0.33	0.05	0.33	0.02
π_2	0.33	0.36	0.13	0.33	0.06	0.33	0.02
π_3	0.34	0.32	0.08	0.34	0.04	0.34	0.02

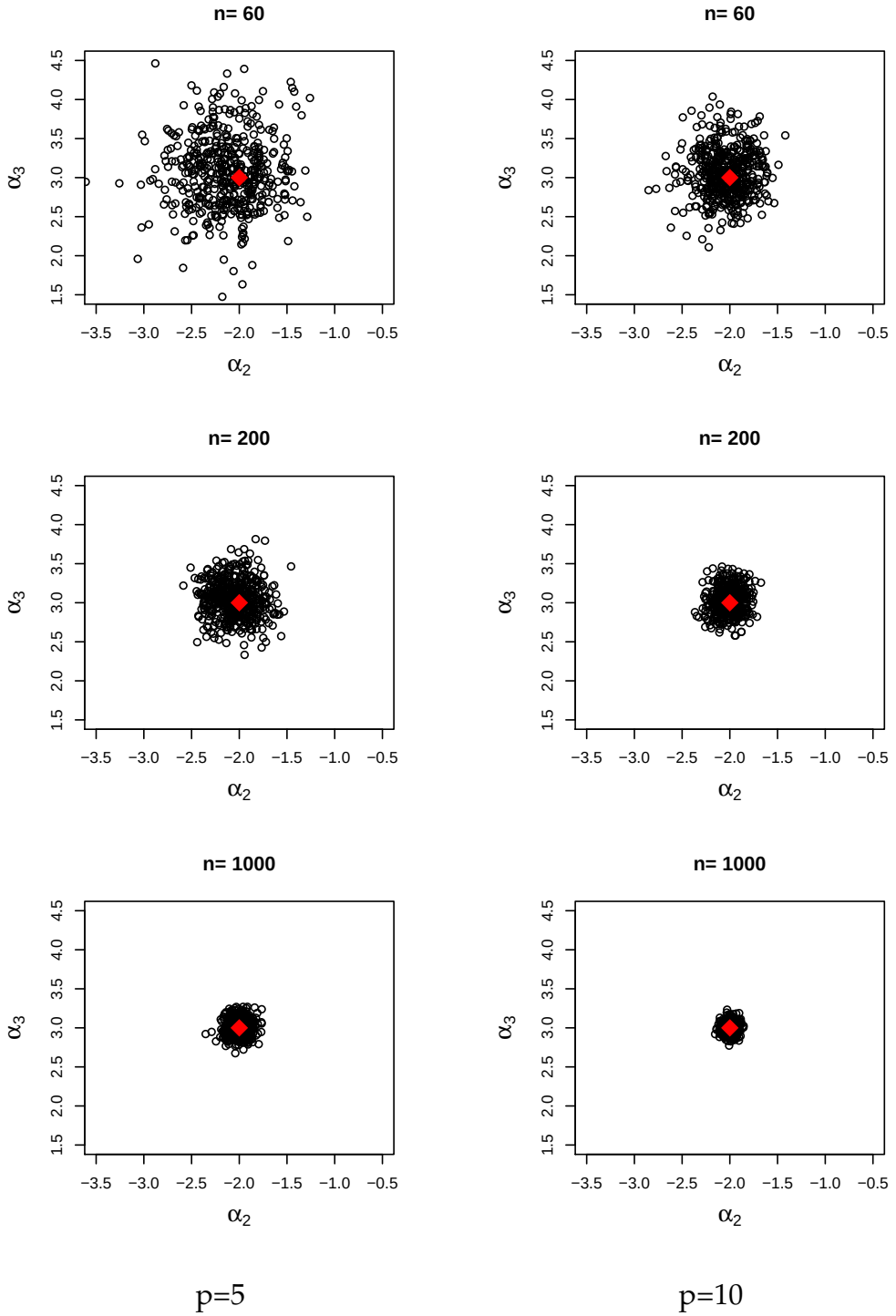


Figure 3.1: Scenarios 1-6. Estimates for (α_2, α_3) in 500 simulated datasets for the POM with row-clustering and column effects. Diamond at the center represents true values $(-2, 3)$.

Table 3.5: Scenarios 10-12. Estimated parameters for the POM with row-clustering and column effects. Mean and standard deviation (sd) over 500 simulated datasets. $\pi = (0.33, 0.33, 0.34)$ and $p = 10$.

		$n = 60$		$n = 200$		$n = 1000$	
Param	true	mean	sd	mean	sd	mean	sd
μ_1	-2.08	-2.14	0.36	-2.10	0.18	-2.09	0.08
μ_2	-1.39	-1.43	0.36	-1.41	0.17	-1.40	0.08
μ_3	1.39	1.39	0.33	1.39	0.16	1.38	0.07
μ_4	2.08	2.10	0.33	2.08	0.17	2.08	0.08
α_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
α_2	-2.00	-2.06	0.25	-2.02	0.13	-2.01	0.06
α_3	3.00	3.06	0.28	3.00	0.14	3.00	0.06
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_2	0.15	0.16	0.36	0.15	0.21	0.15	0.09
β_3	0.30	0.30	0.37	0.30	0.20	0.30	0.09
β_4	0.45	0.45	0.36	0.46	0.20	0.45	0.09
β_5	0.60	0.63	0.37	0.60	0.20	0.60	0.09
β_6	0.75	0.77	0.38	0.76	0.20	0.74	0.09
β_7	0.90	0.92	0.39	0.90	0.22	0.89	0.10
β_8	1.05	1.10	0.39	1.06	0.21	1.05	0.09
β_9	1.20	1.24	0.39	1.21	0.21	1.20	0.10
β_{10}	1.35	1.40	0.38	1.36	0.20	1.35	0.09
π_1	0.33	0.33	0.07	0.33	0.04	0.33	0.02
π_2	0.33	0.34	0.07	0.33	0.04	0.33	0.02
π_3	0.34	0.34	0.06	0.34	0.03	0.34	0.02

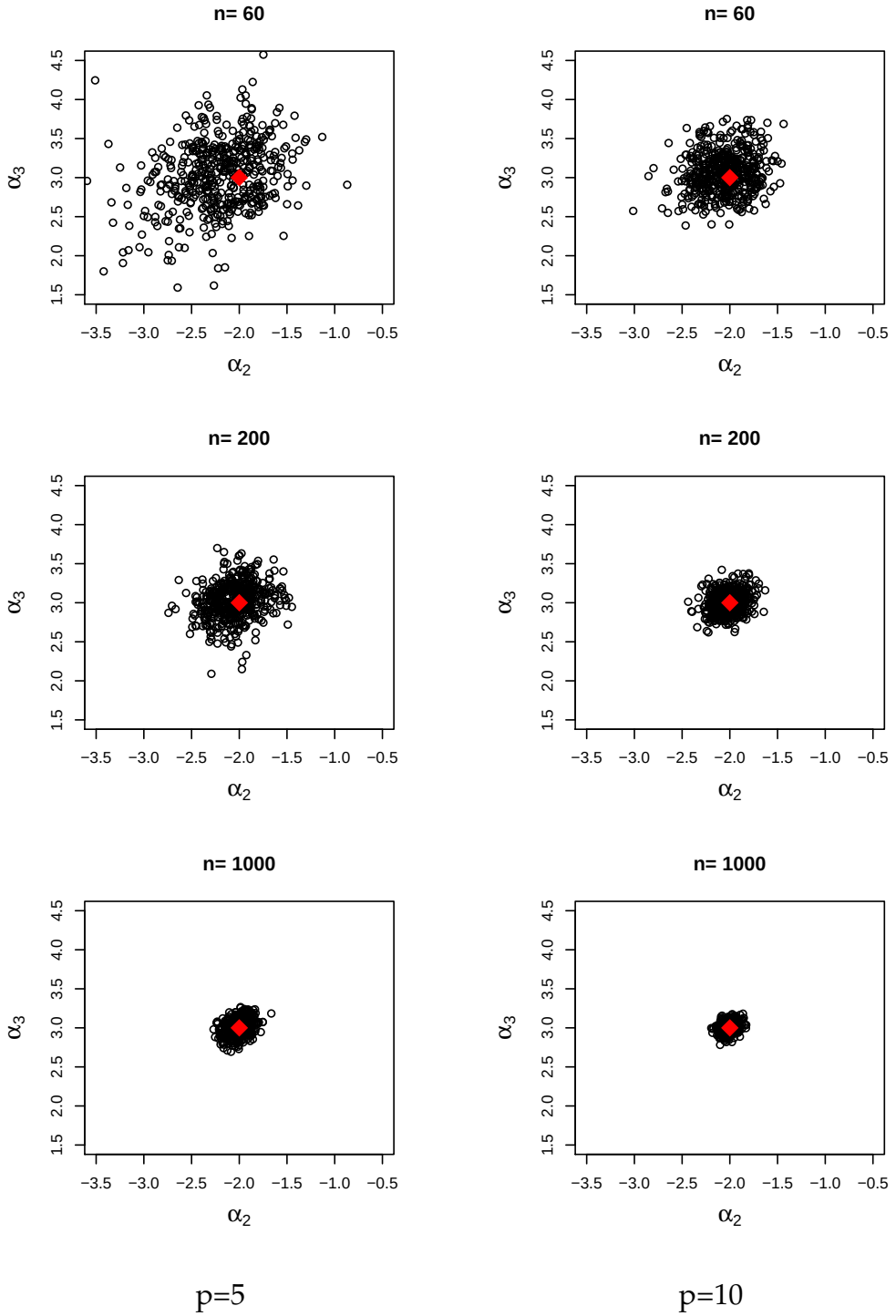


Figure 3.2: Scenarios 7-12. Estimates for (α_2, α_3) in 500 simulated datasets for the POM with row-clustering and column effects. Diamond at the center represents true values $((-2, 3))$.

Next, Tables 3.6 and 3.7 and Figure 3.3 show the results for scenarios 13-18. These scenarios share a very unbalanced mixture proportions $\pi = (0.90, 0.08, 0.02)$ where the smallest component is tiny, only 2% of the total. Not surprisingly, the estimates for all parameters converge to their true values but the variability of the estimated cluster effects α is a lot bigger than in the previous scenarios. For instance, when $n = 60$ and $p = 5$ the standard deviations for α_2 and α_3 are 1.47 and 3.76 (Table 3.6) which are extremely large in comparison to the other scenarios. Doubling the number of columns, $p = 10$ in Table 3.7, reduces these sd's to 0.54 and 2.99 but these are still pretty large. Only when $n = 1000$ the estimated mean values for α are close to the true ones (Figure 3.3). Notably, the unbalanced mixture proportions do not seem to affect the estimates for the column effects in the same way, i.e. the mean estimates for β are close to their true values than those for α even for $n = 60$.

Finally, we present 95% coverage rates for α and β for all scenarios in Table 3.8 and Figure 3.4. We define the 95% coverage rate as the proportion of times in the simulations that the true value is included in the 95% confidence interval for the estimated parameter. On one hand, the coverage rates for the column effect β remain mostly unaffected by either sample size or mixture proportions. The coverage rates for the cluster effects α in Scenarios 1-12, where the mixtures proportions are not highly unbalanced, increase with sample size and are about 80-90% when $p = 5$. In contrast to that, when the mixture proportions are highly unbalanced in Scenarios 13-18 the coverages rates drop drastically to around 50-70%. In particular, the smallest mixture component has the lowest coverages rate. Increasing the number of columns does not change this much, in general 95% coverage rates in Scenarios 1-12 are higher but lower than the ones in Scenarios 13-18.

In conclusion, in order to obtain a meaningful estimates in mixture models with very unbalanced proportions it is crucial to use large sample sizes. Correspondingly care should be taken when mixtures are estimated

using small datasets.

Table 3.6: Scenarios 13-15. Estimated parameters for the POM with row-clustering and column effects. Mean and standard deviation (sd) over 500 simulated datasets. $\pi = (0.90, 0.08, 0.02)$ and $p = 5$.

		$n = 60$		$n = 200$		$n = 1000$	
Param	true	mean	sd	mean	sd	mean	sd
μ_1	-2.08	-2.11	0.48	-2.09	0.23	-2.08	0.08
μ_2	-1.39	-1.38	0.47	-1.39	0.22	-1.39	0.07
μ_3	1.39	1.46	0.46	1.41	0.22	1.39	0.07
μ_4	2.08	2.16	0.47	2.11	0.23	2.09	0.08
α_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
α_2	-2.00	-1.98	1.47	-2.11	0.62	-2.02	0.18
α_3	3.00	2.96	3.76	3.47	2.40	3.03	0.50
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_2	0.15	0.18	0.36	0.15	0.20	0.15	0.09
β_3	0.30	0.32	0.37	0.30	0.21	0.30	0.09
β_4	0.45	0.47	0.35	0.46	0.19	0.45	0.08
β_5	0.60	0.63	0.36	0.61	0.19	0.60	0.09
π_1	0.90	0.66	0.30	0.85	0.15	0.90	0.02
π_2	0.08	0.18	0.22	0.10	0.10	0.08	0.02
π_3	0.02	0.16	0.25	0.05	0.12	0.02	0.01

3.4 Model selection

Model selection criteria for finite mixtures is an active area of research in Statistics with no theoretical foundation completely developed to date. In this section, we thus rely on guidance from simulation studies (McLachlan & Peel 2000, Fonseca & Cardoso 2007, Cubaynes et al. 2012, Fernández & Pledger 2015). We use three Frequentist measures: the Akaike Infor-

Table 3.7: Scenarios 13-15. Estimated parameters for the POM with row-clustering and column effects. Mean and standard deviation (sd) over 500 simulated datasets. $\pi = (0.90, 0.08, 0.02)$ and $p = 10$.

		$n = 60$		$n = 200$		$n = 1000$	
Param	true	mean	sd	mean	sd	mean	sd
μ_1	-2.08	-2.15	0.53	-2.11	0.20	-2.08	0.07
μ_2	-1.39	-1.44	0.52	-1.41	0.20	-1.39	0.07
μ_3	1.39	1.39	0.54	1.39	0.20	1.39	0.06
μ_4	2.08	2.09	0.54	2.08	0.20	2.08	0.07
α_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
α_2	-2.00	-1.93	0.54	-2.01	0.23	-2.01	0.09
α_3	3.00	2.81	3.05	3.24	1.57	3.01	0.21
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_2	0.15	0.14	0.35	0.15	0.20	0.15	0.09
β_3	0.30	0.29	0.36	0.30	0.19	0.30	0.09
β_4	0.45	0.45	0.34	0.46	0.20	0.45	0.09
β_5	0.60	0.62	0.35	0.60	0.19	0.60	0.09
β_6	0.75	0.76	0.33	0.75	0.19	0.75	0.09
β_7	0.90	0.91	0.34	0.90	0.19	0.90	0.09
β_8	1.05	1.09	0.36	1.06	0.19	1.05	0.08
β_9	1.20	1.23	0.36	1.21	0.19	1.21	0.08
β_{10}	1.35	1.37	0.35	1.35	0.19	1.35	0.09
π_1	0.90	0.75	0.30	0.89	0.06	0.90	0.01
π_2	0.08	0.15	0.21	0.09	0.06	0.08	0.01
π_3	0.02	0.10	0.23	0.02	0.01	0.02	0.00

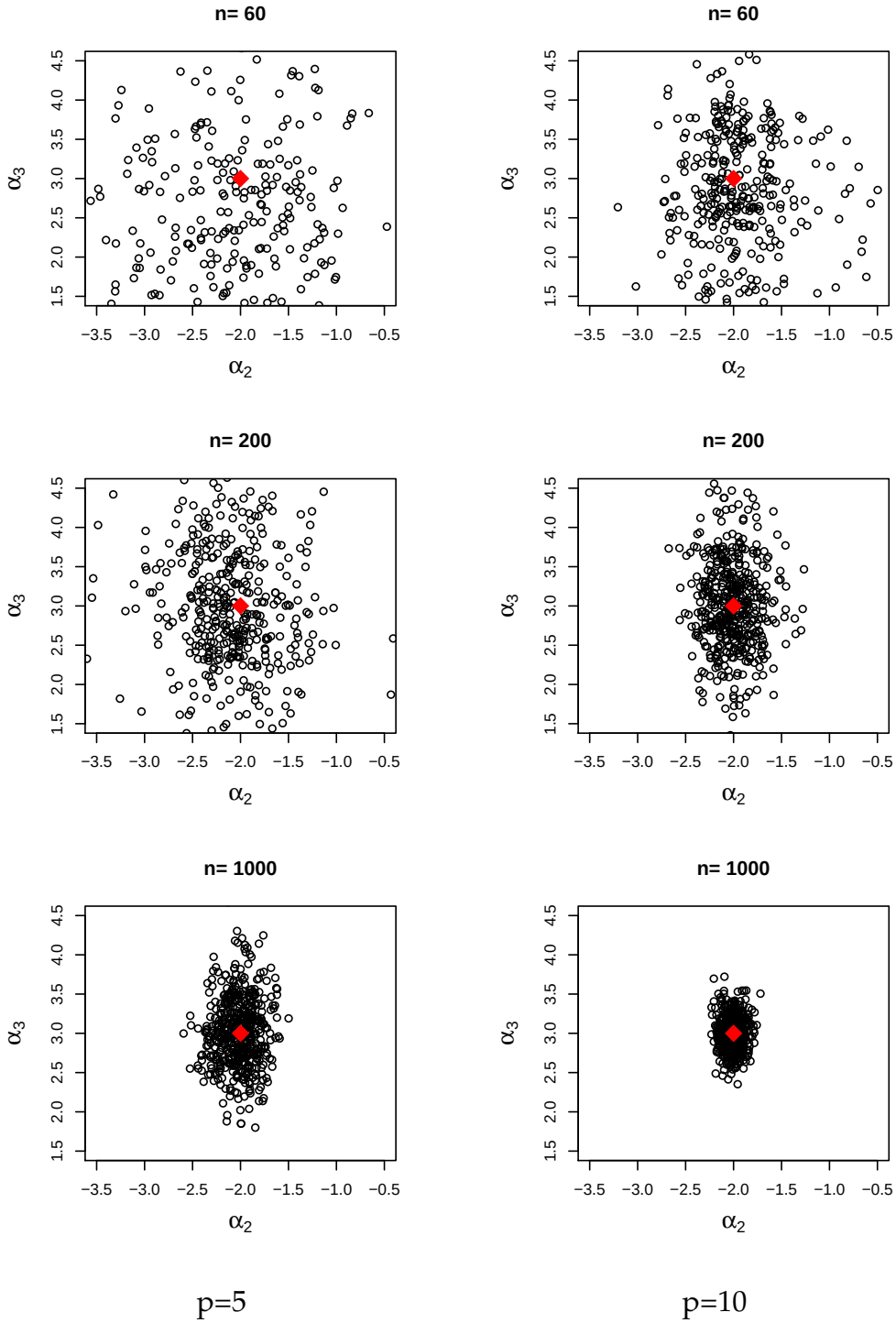


Figure 3.3: Scenarios 13-18. Estimates for (α_2, α_3) in 500 simulated datasets for the POM with row-clustering and column effects. Diamond at the center represents true values $((-2, 3))$.

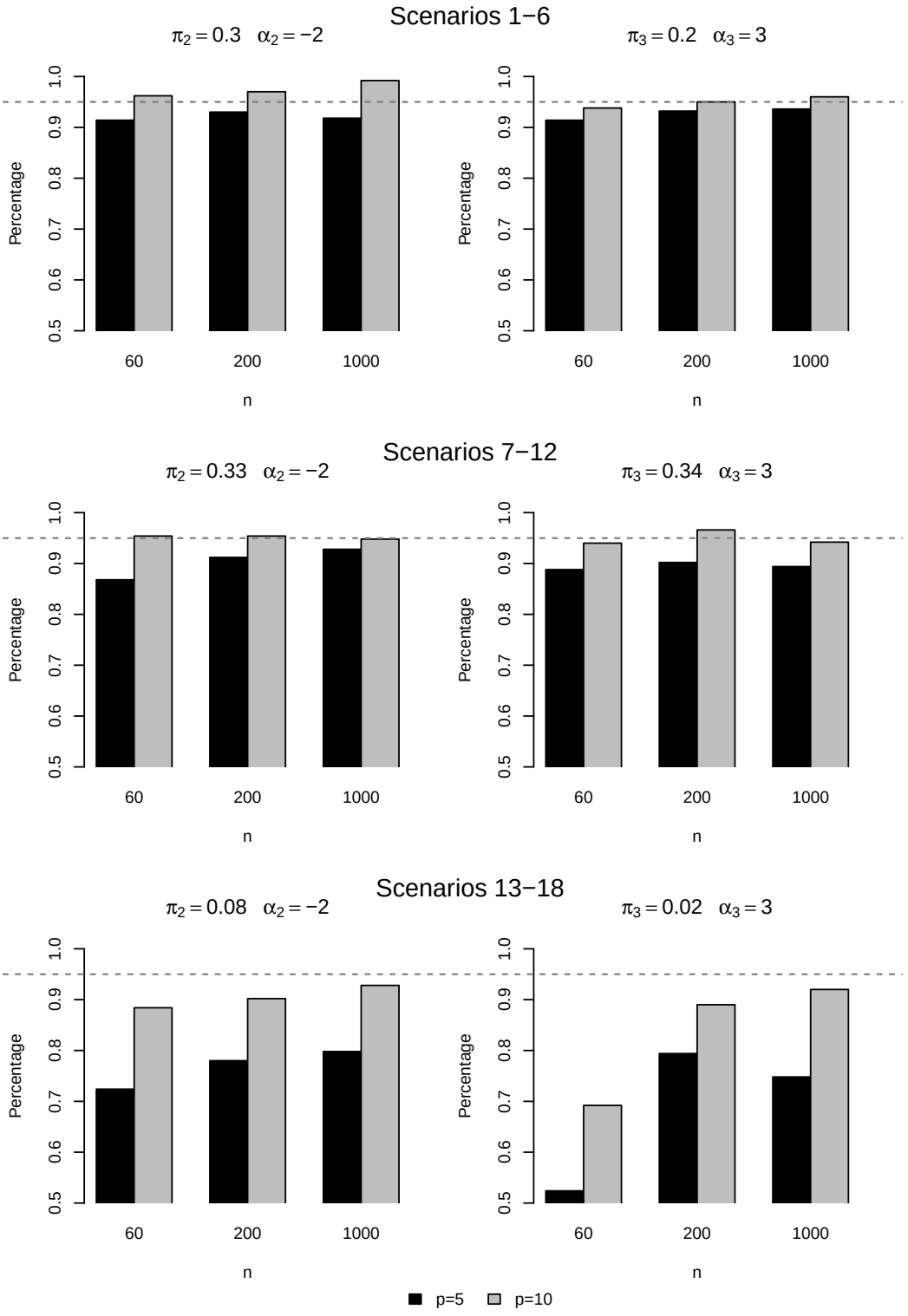


Figure 3.4: 95% Coverage rates for α_2 and α_3 the POM with row-clustering and column effects. All scenarios.

Table 3.8: 95% Coverage rates for the POM with row-clustering and column effects. All scenarios.

		Scenarios 1-3 $\pi = (0.50, 0.30, 0.20)$			Scenarios 7-9 $\pi = (0.33, 0.33, 0.34)$			Scenarios 13-15 $\pi = (0.90, 0.08, 0.02)$		
		n			n			n		
	Param	60	200	1000	60	200	1000	60	200	1000
p=5	α_2	0.91	0.93	0.91	0.87	0.91	0.91	0.72	0.78	0.80
	α_3	0.91	0.93	0.93	0.88	0.9	0.89	0.52	0.79	0.75
	β_2	0.95	0.94	0.95	0.94	0.94	0.96	0.95	0.94	0.95
	β_3	0.93	0.93	0.94	0.95	0.93	0.94	0.95	0.94	0.94
	β_4	0.95	0.95	0.96	0.96	0.95	0.96	0.96	0.95	0.97
	β_5	0.93	0.94	0.96	0.95	0.94	0.95	0.96	0.96	0.95
		Scenarios 4-6 $\pi = (0.50, 0.30, 0.20)$			Scenarios 10-12 $\pi = (0.33, 0.33, 0.34)$			Scenarios 16-18 $\pi = (0.90, 0.08, 0.02)$		
		n			n			n		
	Param	60	200	1000	60	200	1000	60	200	1000
p=10	α_2	0.96	0.97	0.98	0.95	0.95	0.93	0.88	0.90	0.93
	α_3	0.94	0.95	0.96	0.94	0.97	0.94	0.69	0.89	0.92
	β_2	0.96	0.92	0.96	0.96	0.93	0.96	0.96	0.95	0.95
	β_3	0.96	0.93	0.96	0.95	0.97	0.97	0.97	0.96	0.96
	β_4	0.95	0.96	0.95	0.96	0.96	0.97	0.97	0.95	0.95
	β_5	0.94	0.96	0.95	0.95	0.96	0.94	0.96	0.97	0.94
	β_6	0.92	0.94	0.94	0.94	0.95	0.97	0.96	0.96	0.95
	β_7	0.95	0.94	0.96	0.95	0.93	0.93	0.97	0.94	0.96
	β_8	0.95	0.94	0.95	0.94	0.94	0.95	0.94	0.96	0.95
	β_9	0.95	0.95	0.94	0.94	0.96	0.95	0.96	0.96	0.95
	β_{10}	0.96	0.95	0.94	0.94	0.96	0.96	0.96	0.96	0.95

mation Criterion (AIC), Akaike (1973), the Bayesian Information Criterion (BIC), Schwarz (1978), and the integrated classification criterion (ICL) by Biernacki et al. (2000).

The lack of a unified coherent theoretical foundation in this area has occurred because comparing finite mixture models with different number of clusters violates classical regularity conditions (Cramér 1946) and therefore the use of criteria based on the likelihood is doubtful. In particular, maximum likelihood estimates under the null hypothesis (reduced model) are on the boundary of the parameter space, e.g. the reduced model has a smaller number of mixture components and thus at last one of the mixture proportions is zero. Secondly, the null hypothesis corresponds to a non-identifiable subset of the parameter space (McLachlan & Peel 2000, section 6.4). This occurs because a mixture with g components could also be re-expressed as a mixture of $g + 1$ components, for example by doubling up one of its components and halving their mixture probabilities. As a result of these two violations, the asymptotic distribution of the test statistic under the null hypothesis is not the usual χ^2 with degrees of freedom equal to the difference in the number of parameters under both hypothesis. In general, this asymptotic distribution under the non-identifiable case is unknown. To date, there are conjectures and simulations for some cases, in particular for continuous data, but not for mixtures of ordinal data.

The AIC and BIC are defined in terms of the maximised incomplete data log-likelihood ℓ and a penalty for model complexity:

$$AIC = -2\ell(\hat{\Omega}, Y) + 2v$$
$$BIC = -2\ell(\hat{\Omega}, Y) + v\log(np)$$

where: $\hat{\Omega}$ are the model parameters estimated by ML, v the number of model parameters, n= number of rows, and p = number of columns. Lower values of the AIC and BIC are an indication of better fit. Both criteria differ only in the second term with the BIC having a penalty that depends on the sample size. This penalty is higher than the one in AIC whenever $\log(np) > 2$.

Simulation studies have shown that AIC and BIC tend to overestimate the number of mixture components for continuous data (McLachlan & Peel 2000, Cubaynes et al. 2012). However, this might not be the case when modelling other types of data. Recently, for example (Fonseca & Cardoso 2007, Fernández & Pledger 2015) used simulations of categorical and ordinal data and found that BIC correctly identifies the number of mixture components in wide range of scenarios.

The ICL is classification-based information criteria that also takes into account the degree of separation of the estimated mixture components, that is the fuzzyness of the estimated clusters. The ICL is formed from the maximised complete data likehood ℓ_c (for example 3.25 in the bi-clustering case) and a term to take into account the fuzziness of the estimated clusters. This term is also known as *entropy* and acts as a penalty for the degree of separation of the mixture components. Models with well separated mixture components will have small entropy whereas poor separation in the mixture components will lead to large entropy. As a result, the ICL takes into account both the model's complexity (number of parameters) and how fuzzy the cluster allocation is. Here we use the large-sample approximation for the ICL, the ICL-BIC (Biernacki et al. 2000) calculated as

$$ICL - BIC = -2\ell_c(\hat{\Omega}, Y) + v\log(np)$$

where: ℓ_c is the complete information log-likelihood.

As can be seen, the BIC and the ICL-BIC differ only in their first term. It has been shown to correctly selected the number of clusters in simulations when the mixing proportions are equal (McLachlan & Peel 2000, Biernacki et al. 2000). It has a similar behaviour to BIC but it does not require the evaluation of the incomplete information likelihood L. This evaluation is computationally very expensive, especially for large datasets and when dealing with bi-clustering. In this latter case, it requires the consideration of either all possible combinations of the p columns to C groups or all

possible combinations of the n rows to the R groups, refer to (3.17) and (3.18).

3.5 Application: 2009-2013 Life Satisfaction in New Zealand

In this section we use self-reported "Life Satisfaction" (LS) from the New Zealand Attitudes and Values survey (NZAVS) to illustrate the models presented so far. We use all individuals with complete responses between 2009 and 2013, and thus this dataset has 2564 rows (n), 5 columns (p) and 7 ordinal levels (q). A detail description of this dataset and the NZAVS survey could be found in Chapter 2.

We fit a wide range of different models (51) to this data and use the AIC, BIC and ICL-BIC to compare amongst models. Specifically, we estimate the row, column and bi-cluster models, section 3.1, using $R = 2 \dots 10$ and $C = 2, 3$. In order to check for potential time trends in LS we also include column effects (year) and interactions in the row-clustering case. For completeness, we also include the null, row effects, column effects, and full main effects (row and column effects) models. The results are shown in Table 3.9.

The information criteria provide different answers (highlighted rows in table 3.9). AIC chooses the full main effects model with 2573 parameters, BIC a six-component row-clustered model with column effects with 20 parameters and ICL-BIC a four-component row-clustered model with column effects with 16 parameters. This is not surprising since as seen in the previous section AIC, BIC and ICL-BIC use different penalties for the number of parameters, with AIC penalising complexity the least and ICL-BIC the most. As a balance of parsimony and model complexity we decide to use the model selected by the BIC, that is a model with six row-clusters and a fixed year effect (BIC equal to 31255).

Table 3.9: Model comparison: Life Satisfaction in NZ using the NZAVS

Model	Linear Predictor	R	C	Pars	AIC	BIC	ICL-BIC
Null	μ_k	1	1	6	38205	38307	-
Row effects	$\mu_k - \alpha_i$	2564	1	2569	27551	71769	-
Column effects	$\mu_k - \beta_j$	1	5	10	38196	38366	-
Row and column	$\mu_k - \alpha_i - \beta_j$	2564	5	2573	**27493**	71779	-
Row-clustering	$\mu_k - \alpha_r$	2	1	8	33751	33811	34377
		3	1	10	32031	32106	33047
		4	1	12	31326	31415	32735
		5	1	14	31203	31308	32769
		6	1	16	31148	31268	33249
		7	1	18	31153	31287	33849
		8	1	20	31156	31305	34451
		9	1	22	31125	31290	34925
		10	1	24	31131	31310	35486
Row-clustering+	$\mu_k - \alpha_r - \beta_j$	2	5	12	33727	33816	34380
column effects		3	5	14	31998	32102	33040
		4	5	16	31285	31405	**32717**
		5	5	18	31161	31296	32748
		6	5	20	31105	**31255**	33243
		7	5	22	31109	31274	33903
		8	5	24	31081	31261	34567
		9	5	26	31074	31268	34709
		10	5	28	31087	31297	35441
Row-clustering+	$\mu_k - \alpha_r - \beta_j - \gamma_{rj}$	2	5	16	33725	33845	34402
column effects+		3	5	22	31996	32160	33092
interactions		4	5	28	31286	31495	32794
		5	5	34	31165	31419	32885
		6	5	40	31107	31406	33409
		7	5	46	31075	31419	33631
		8	5	52	31032	31421	34133
		9	5	58	31027	31460	34639
		10	5	64	31006	31484	34722
Column-clustering	$\mu_k - \beta_c$	1	2	8	38209	38269	38269
		1	3	10	38213	38288	38288

Continued on next page...

Table 3.9 – continued from previous page

Model	Linear Predictor	Rows	Cols	Pars	AIC	BIC	ICL-BIC
Bi-clustering	$\mu_k - \alpha_r - \beta_c$	2	2	10	33755	33830	34396
		2	3	12	33759	33849	34415
		3	2	12	32035	32125	33066
		3	3	14	32039	32144	33085
		4	2	14	31329	31434	32754
		4	3	16	31333	31453	32773
		5	2	16	31207	31327	32787
		5	3	18	31211	31346	32806
		6	2	18	31152	31287	33277
		6	3	20	31156	31306	33295
		7	2	20	31156	31306	33955
		7	3	22	31160	31325	33973
		8	2	22	31127	31291	34419
		8	3	24	31131	31310	34438
		9	2	24	31123	31302	34990
		9	3	26	31127	31321	35009
		10	2	26	31135	31329	36172
		10	3	28	31139	31348	36191

With respect to the trends over time, it is important to notice there is evidence of time variation in LS since the selected model includes year effects (β_j). However, these year effects do not seem to vary by cluster as the incorporation of year-cluster interactions do not improve model fit with BIC increasing from 31255 to 31406 when introducing these interactions to the model. In addition to that, there is no evidence that these year effects can be grouped over time as the column and bi-clustered models do not improve model fit.

As a convention, we name the clusters according to their levels of the $\hat{\alpha}_r : \alpha_1 \dots \alpha_6$ so that the respondents in cluster 1 tend to have the lowest levels of LS and those in cluster 6 tend to have the highest. We then proceed to assign individuals to each estimated cluster, a procedure known as classification in the machine learning literature (Hastie et al. 2009). Specifically, our model-based clustering approach would be called *unsupervised*

classification due to the use of latent/unobserved covariates in the model. Due to the use of mixture models the allocation of individuals to clusters is fuzzy, that is each individual has always a probability of coming from each cluster ($z_{ir} \geq 0$, $\sum_{r=1}^{R} z_{ir} = 1, \forall i$). Allocation r_i of individual i to cluster r is based on a highest a posteriori probability criterion:

$$\hat{r}_i = \underset{r \in 1,\ldots,R}{\operatorname{argmax}} \hat{z}_{ir} \ , \ i = 1 \ldots n \tag{3.26}$$

Where: $\hat{z}_{ir}$ is the posterior probability that individual i belongs to cluster r. See the E-step of the EM algorithm in (3.20).

Visualisation of the estimated clusters using heatmaps

We next use heatmaps to visually assess the fuzziness of this classification of individuals into each of the six groups of the model selected by BIC. It is important to mention that heatmaps should only be used to visualise the best fitting model(s). That is, after statistical estimation of several models and not to compare among all candidate models as the human eye tends to see patterns in any given image (Wilkinson & Friendly 2009).

Estimated allocations $\hat{r}_i$ close to 1 would mean that our fuzzy probabilistic clustering is "crisp". To do so, we calculate the co-clustering probabilities for all individuals. A co-clustering probability is the probability that any pair of individuals (i, i') come from the same cluster r conditional on the model parameters Ω and the observed responses Y. It is defined as follows:

$$\begin{aligned}
C_{ii'} &= \sum_{r=1}^{R} P(z_{ir} = 1, z_{i'r} = 1 | \hat{\Omega}, Y) \\
C_{ii'} &= \sum_{r=1}^{R} P(z_{ir} = 1 | \hat{\Omega}, Y) P(z_{i'r} = 1 | \hat{\Omega}, Y) \\
C_{ii'} &= \sum_{r=1}^{R} \hat{z}_{ir} \hat{z}_{i'r}, i, i' = 1, \ldots n
\end{aligned}$$

Or in matrix form

$$C = ZZ'$$

where: Z is the estimated membership matrix ($n \times R$) in (3.20).

Figure 3.5 shows the co-clustering probabilities for all the respondents in the original and clustered data. As we can see, there is no apparent pattern when we look at the co-clustering probabilities in the original data (upper panel). This changes markedly when we allocate respondents to estimated clusters as the co-clustering probabilities within each cluster are very high, around 0.8 on average. The allocations of individuals to clusters of the selected model ($R = 6$) is therefore crisp. Plotting these co-clustering probabilities also allows us to appreciate the relative size of each cluster. Clusters 1 and 6 are the smallest ones, as they capture respondents at both extremes of the ordinal scale, whereas cluster 3 and 4 are the biggest ones. We will see this more in detail when looking at the estimated cluster proportions ($\hat{\pi}_r$ in Figure 3.7).

What do these estimated six row-clusters selected by the BIC look like? The estimated clusters are shown in Figures 3.6 and 3.7. Heatmaps in Figure 3.6 offer a first overall look. This figure shows both the original (upper panel) and clustered data (lower) with rows, columns and cell colours representing individuals, occasions and ordinal levels (red=1 to white=7). For instance, the cluster at the bottom of the lower panel is very small and mostly red. It is thus is composed by individuals whose life satisfaction is at lower end of the ordinal scale. In contrast to that, a bigger cluster near the middle is mostly white and therefore has individuals whose responses are at the upper end.

Next, Figure 3.7 displays the distribution of LS by cluster and year. It also shows for each cluster the estimated proportions, $\hat{\pi}_r$, and cluster effects $\hat{\alpha}_r$. In this Figure, we can clearly see that all the clusters have different LS response patterns. For example, cluster 1 is about 1% of the all the sample and is composed by individuals that have extremely low levels of LS.

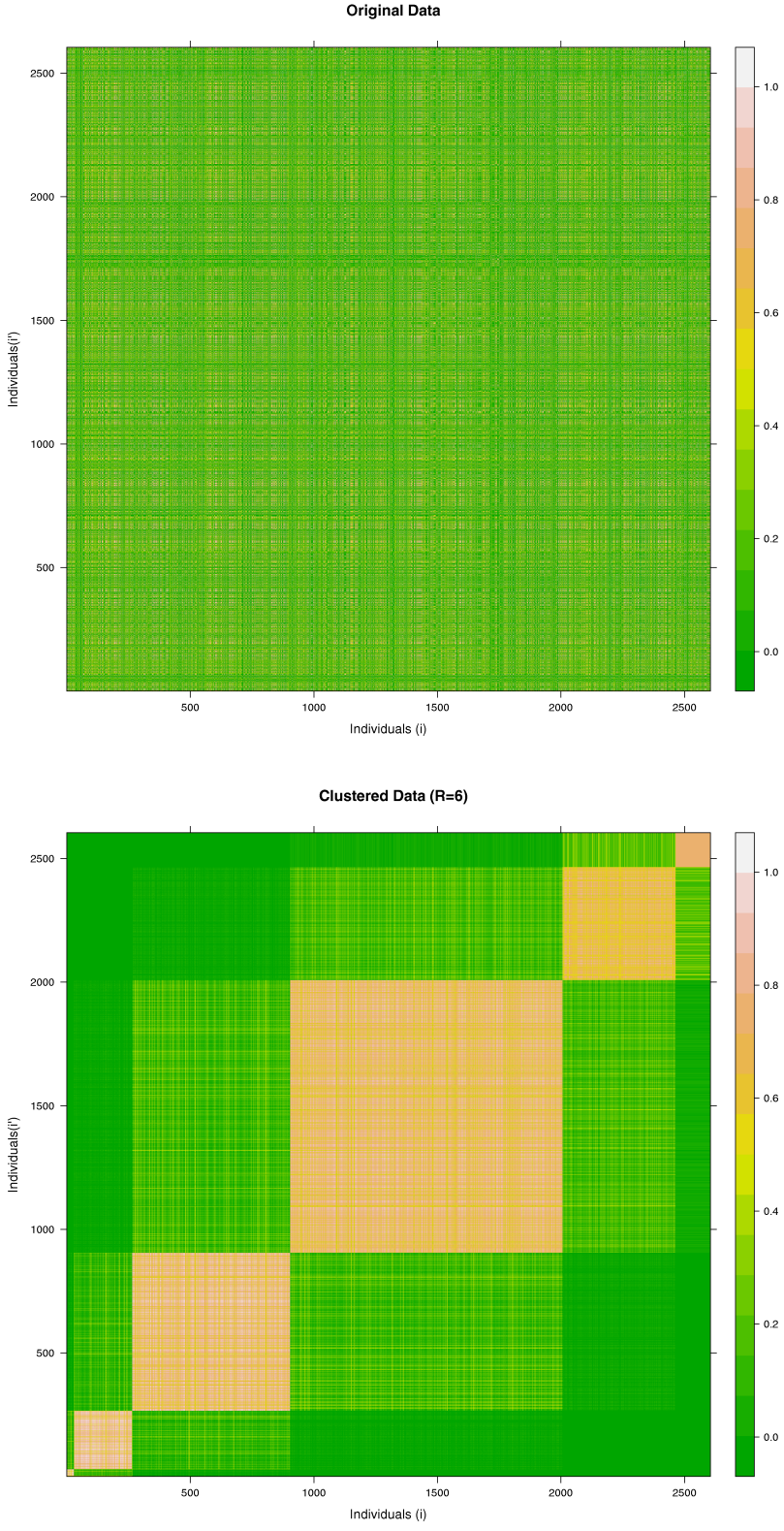


Figure 3.5: Co-clustering probabilities $C_{ii'}$ for all respondents in the original and clustered data

They are all very close to 1. Similarly, individuals in cluster 2 are around 9% of the total and have a more neutral view of their life satisfaction with levels closer to the middle of the ordinal scale ($\hat{\alpha}_2 = 2.8$). In contrast to that, clusters 5 and 6 are formed by people that are extremely satisfied with their life over 2009-2013 ($\hat{\alpha}_5 = 10.2$ and $\hat{\alpha}_5 = 12.8$). Together, both groups are around a quarter of the whole sample ($\pi_5 = 18\%$ and $\pi_6 = 6\%$). Finally, clusters 3 and 4 are the ones whose LS patterns over time resemble the most the overall population. Unsurprisingly, they are also the ones with the highest proportions ($\pi_3 = 25\%$, and $\pi_4 = 40\%$).

Comparison of estimated clusters with socio-economic variables in the NZAVS

LS is known to be correlated with several socio-economic factors such as employment, ethnicity, qualifications and many others. This begs the question: how do our estimated clusters compare to these demographics? Put it simply, which observables have been captured by our latent variable approach?. Given that NZAVS is a household survey with plenty of socio-economic information we can attempt to answer these questions. As starting point here, we compare the estimated clusters with socio-economic deprivation and income. We use the New Zealand index of socio-economic deprivation in 2013 (NZDep2013) by Atkinson et al. (2014) and total household income as proxies for socio-economic deprivation and income.

The NZDep2013 is an area-based measure that estimates the level of deprivation for people in a defined small area, or meshblock. Statistics New Zealand defines a meshblock as a geographical area with a population of around 60 to 110 people. The NZDep 2013 is based on nine Census variables: income, home ownership, employment, qualifications, family structure, housing, access to transport and communications. Originally coded in deciles (1-10), we transform NZDep2013 into quintiles (Q1-Q5) so that respondents in Q5 live in the most deprived areas of the country.

Figure 3.8 shows boxplots for the income distribution by NZDep2013

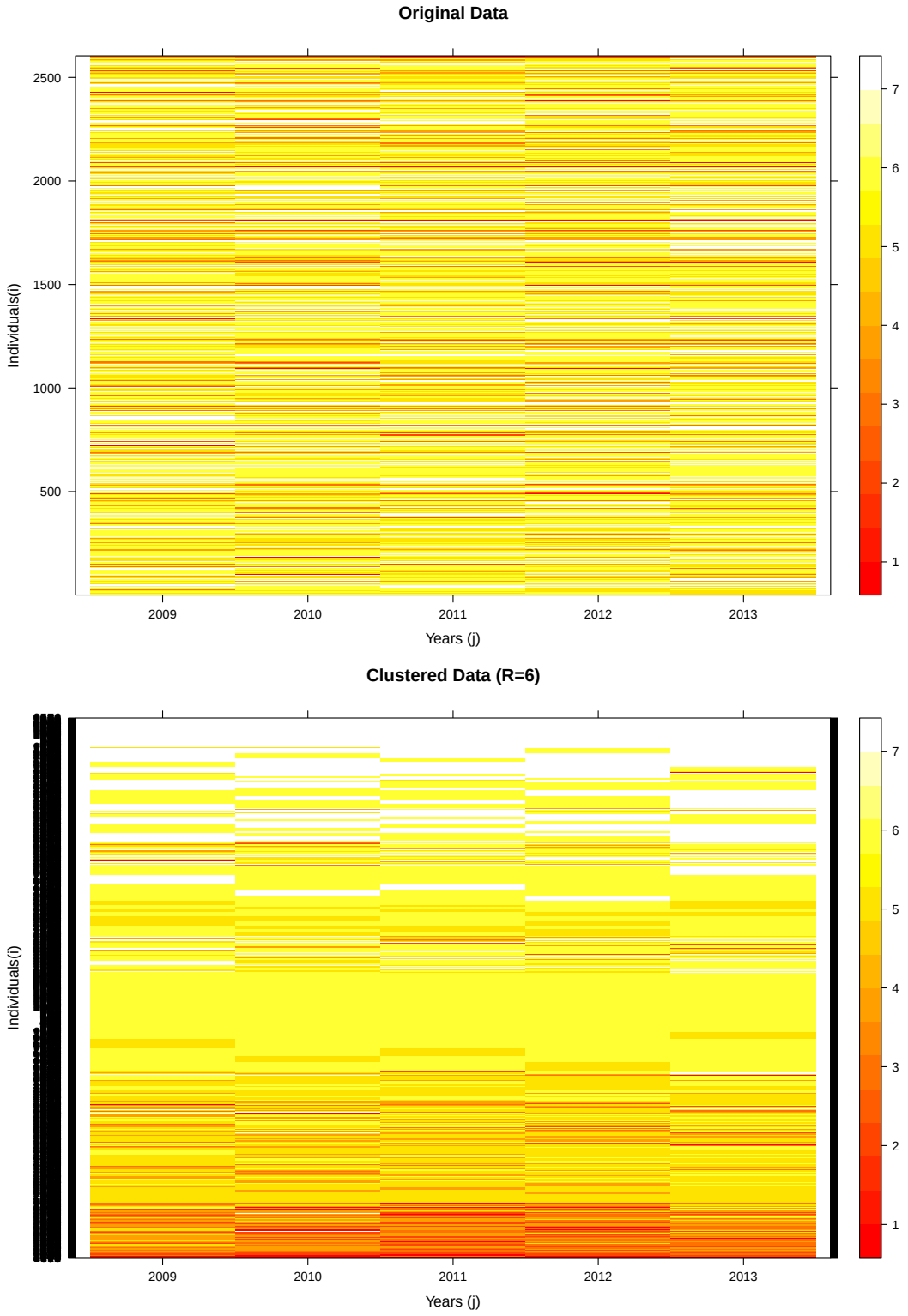


Figure 3.6: Heatmaps for life satisfaction data ($y_{ij} = k$). Rows, columns and cell colours represent individuals (i), years (j) and ordinal levels (k =Strongly Disagree (red) ... Strongly Agree (white)).

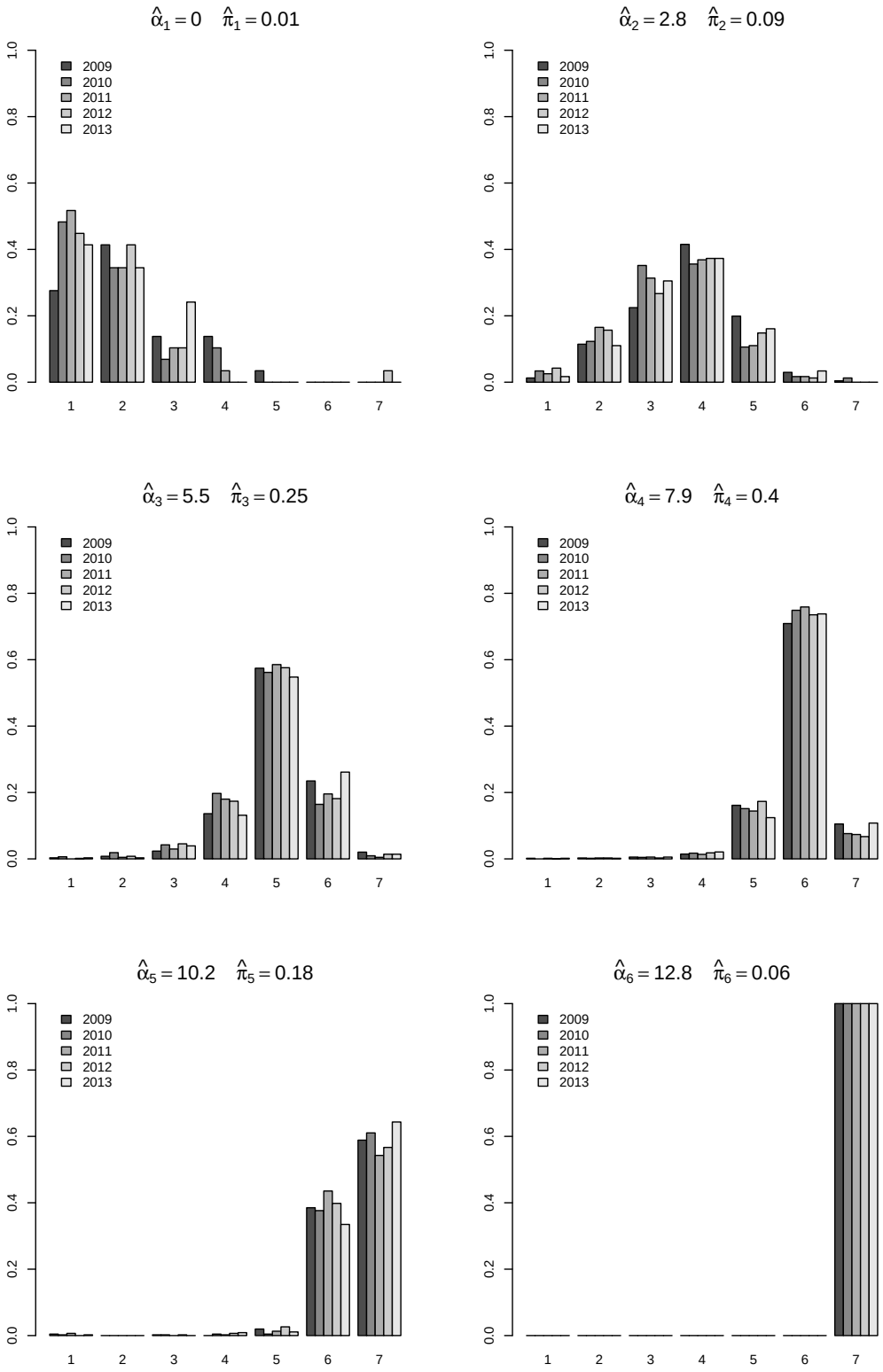


Figure 3.7: Distribution of life Satisfaction in the NZAVS by cluster and year

quintile and cluster. When looking at the income distribution by quintile (left) when can see that income monotonically decreases with deprivation. Although there is a substantial amount of overlap we could see that on average more deprived people have less income. Furthermore, the median income of people in the middle quintile (Q3) equates the overall median income (NZ80,000) . Now, looking at the income distribution by cluster (right) we can see that the income distribution varies greatly by cluster. In general, the estimated cluster effect is positively correlated with the median income in the cluster. Individuals in clusters 1 and 2 for example have the lowest cluster effects ($\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 = 2.8$) and thus the lowest levels of LS (Figure 3.7). They also have the lowest median incomes which are well below the overall median income . However, this relation is not monotonic. Individuals in clusters 4, 5 and 6 have almost identical median incomes, all above the overall median, but different cluster effects ($\hat{\alpha}_4 = 7.9$, $\hat{\alpha}_5 = 10.2$ and $\hat{\alpha}_6 = 12.8$) and thus have different LS response patterns. Therefore, the estimated latent cluster effects $\hat{\alpha_r}$ are capturing the effects of income but also some other effects.

With regard to socio-economic deprivation, Table 3.10 shows the distribution of NZDep2013 quintiles in the overall NZ population and on each cluster. Here we see that this distribution varies greatly by cluster. In particular, clusters 1 and 2 having the lowest levels of LS also have much higher proportions of more deprived respondents. The proportions of people of clusters 1 and 2 in Q5 are 31% and 24% whereas it is only 12% in the general population. Conversely, cluster 6, formed by people extremely satisfied by life, has a higher proportion of less deprived respondents, 31 and 27% for Q1 and Q2 against 26 and 25% in the general population. Therefore, the estimated cluster effects $\hat{\alpha}$ correlate better with deprivation than income. It is important to notice that, although the estimated clusters have different deprivation patterns, they have respondents from all levels of deprivation. That is, clusters with low/high LS levels are not formed exclusively by the most/least deprived people. For instance,

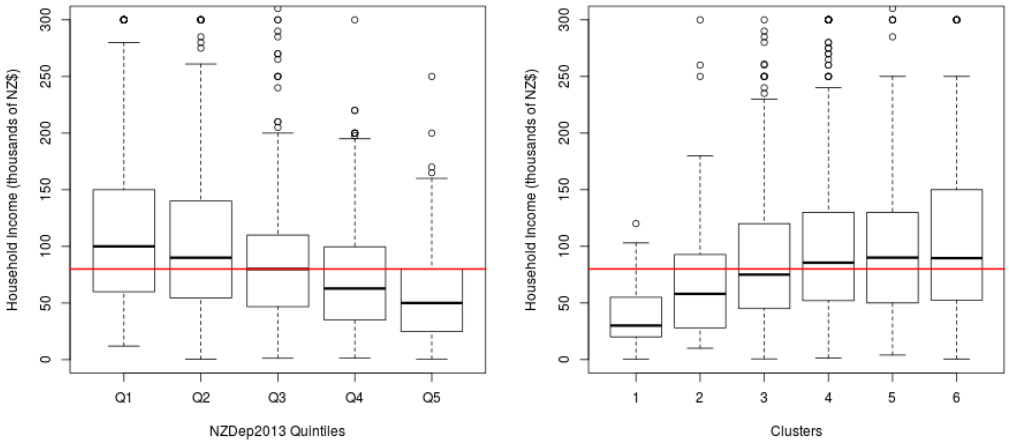


Figure 3.8: Income Distribution by NZDep quintiles (Q1-Q5) and estimated clusters. Solid line represents the overall median income (NZ80,000) for the whole NZAVS sample

in cluster 1, 1% of the total and where respondents are extremely dissatisfied with life, still has 10 and 21% of least deprived quintiles Q1 and Q2. Likewise, cluster 5, which accounts for 18% of the total and where respondents are very positive about their life satisfaction, has 16 and 10 % of people from Q4 and Q5 the most deprived quintiles.

In summary, the model-based clustering methods presented in this chapter allows us to identify groups of individuals with different socio-economic profiles based solely on their LS responses over time. Next, in Chapter 4, although still assuming that responses within individuals are independent, we augment the linear predictor of the clustering models presented so far to incorporate non-proportional odds.

Table 3.10: Proportions of individuals by NZDep quintiles (Q) by cluster and overall

Cluster (r)	$\hat{\pi}_r$	$\hat{\alpha}_r$	Q1	Q2	Q3	Q4	Q5
1	0.01	0	**0.10**	**0.21**	**0.10**	**0.28**	**0.31**
2	0.09	2.8	**0.19**	**0.17**	**0.18**	**0.23**	**0.24**
3	0.25	5.5	0.24	0.23	0.21	0.21	0.11
4	0.40	7.9	0.28	0.27	0.20	0.15	0.10
5	0.18	10.2	0.28	0.27	0.19	0.16	0.10
6	0.06	12.8	**0.31**	**0.27**	**0.16**	**0.15**	**0.11**
		Overall	0.26	0.25	0.19	0.17	0.12

Chapter 4

Trend Odds Model

4.1 Model

In this chapter, we extend the Trend Odds model (TOM) of Capuano & Dawson (2012) to incorporate latent clusters. Similarly to Chapter 3, the models here assume that the observations are independent over time. The TOM is a monotone constrained non-proportional odds models that adds an extra parameter to the linear predictor of the cumulative probability, allowing parsimonious incorporation of non-proportional odds.

As before, the data Y is a (n, p) matrix where each cell y_{ij} is equal to any of the q ordinal categories, where: $i = 1, ..., n$; $j = 1, ..., p$ and $k = 1, ..., q$. We now fully describe the TOM for the row-clustering case and then summarize the main characteristics of the other models in Table 2.2.

Row-clustering

We start with the case of row-clustering. Rows are assumed to come from any of the R row groups with a priori probabilities $\pi_1, \ldots, \pi_R$. That is, we assume that the rows come from a finite mixture with R components where both R and the row-cluster proportions π_r are unknown. Note also that $R < n$ and $\sum_{r=1}^{R} \pi_r = 1$. Let θ_{rjk} be the probability that observation

$y_{ij} = k$ given that row i belongs to row-cluster r. That is $P(y_{ij} = k|i \in r) = \theta_{rjk}$. Given a trend parameter by cluster γ_r and an arbitrary scalar t_k that varies by ordinal outcome k, the TOM with clustering has the form

$$\text{Logit}[P(y_{ij} \leq k|i \in r)] = \mu_k - \alpha_r - \gamma_r t_k \tag{4.1}$$

or equivalently

$$\theta_{rjk} = \frac{\exp(\mu_k - \alpha_r - \gamma_r t_k)}{1 + \exp(\mu_k - \alpha_r - \gamma_r t_k)} - \frac{\exp(\mu_{k-1} - \alpha_r - \gamma_r t_{k-1})}{1 + \exp(\mu_{k-1} - \alpha_r - \gamma_r t_{k-1})} \tag{4.2}$$

where $\alpha_1 = \gamma_1 = 0$ and $\mu_k - \mu_{k-1} \geq \gamma_r(t_k - t_{k-1})$. The parameter μ_k is the k^{th} cut point and α_r is the effect of row-cluster r. The parameter γ_r can also be interpreted a shape parameter for θ_{rjk}. The term $\gamma_r t_k$ represents any non-proportional odds for the cluster since it depends both on r and k. Following Capuano & Dawson (2012), we set to $t_k = k - 1$. Given this, the constraint necessary to make sure the cumulative probabilities are non-decreasing becomes

$$\mu_k - \mu_{k-1} \geq \gamma_r \qquad \forall r, k \tag{4.3}$$

As an example, Figure 4.1 displays the probability distribution of θ_{rjk} for the POM and the TOM with clustering. We use an ordinal response with five outcomes ($q = 5$), three row-clusters ($R = 3$) and the following values for the parameters: $\mu = (-1.95, -1.10, 1.10, 1.95)$, $\alpha = (0, -3, 3)$ and $\gamma = (0, -2, -1)$. Notice that θ_{rjk} have the same shape for all clusters but different location (α_r) in the case of the POM. In contrast to that, for the TOM θ_{rjk} have both different shape (γ_r) and location (α_r) in all clusters. See also Figure 1.3 in Chapter 1 that shows a graphical representation of the original formulation of the TOM.

Assuming independence over the rows and, conditional on the rows, independence over the columns, the likelihood for the TOM with row-

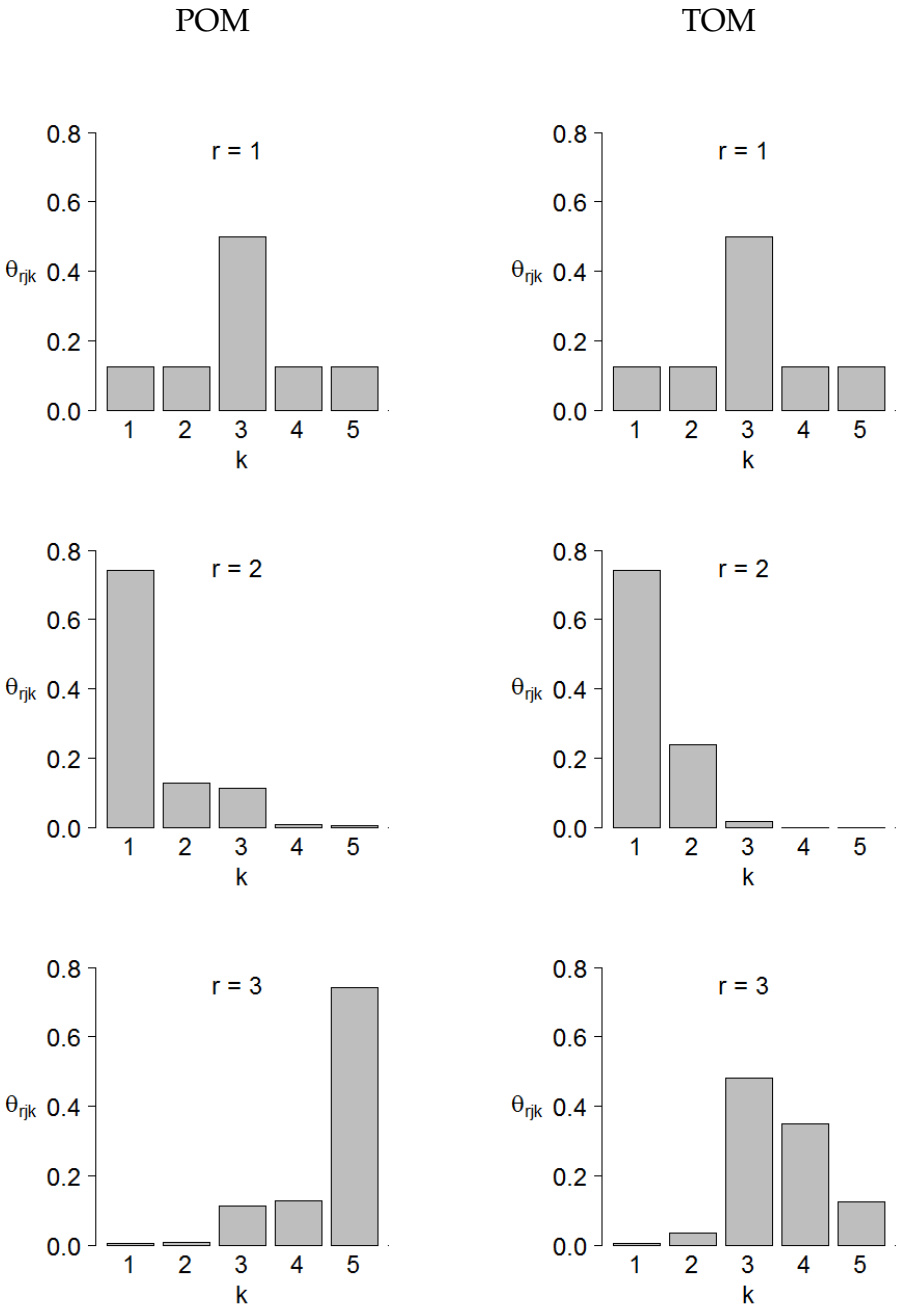


Figure 4.1: Probability distribution for θ_{rjk}

clustering becomes

$$L(\phi, \pi|Y) = \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)} \tag{4.4}$$

where ϕ is the set of model parameters (μ, α, γ). The expression above is also referred as incomplete data likelihood given that the cluster memberships are unknown. The number of model parameters is equal to: $v = (q-1) + 3(R-1)$.

Table 4.1 shows the linear predictor, the constraint for non-decreasing cumulative probabilities and the number of parameters for all the TOM models with clustering.

Table 4.1: Summary of the TOM models with clustering

Model	Linear Predictor ϕ	Additional Constraints	Number of Parameters
Row clustering	$\mu_k - \alpha_r - \gamma_r t_k$	$\mu_k - \mu_{k-1} \geq \gamma_r \ , \ \forall r, k$ $\gamma_1 = 0$	$(q-1) + 3(R-1)$
Column clustering	$\mu_k - \beta_c - \delta_c t_k$	$\mu_k - \mu_{k-1} \geq \delta_c \ , \ \forall c, k$ $\delta_1 = 0$	$(q-1) + 3(C-1)$
Bi-clustering	$\mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c) t_k$	$\mu_k - \mu_{k-1} \geq \gamma_r + \delta_r \ , \ \forall r, c, k$ $\gamma_1 = \delta_1 = 0$	$(q-1) + 3(R+C-2)$

4.2 Model selection

In this chapter, we will use the same information criteria we used for the Proportional Odds model in Chapter 3, namely: AIC (Akaike 1973), BIC (Schwarz 1978), and ICL-BIC (Biernacki et al. 2000).

4.3 Simulations

In order to check that we are able to recover the true model parameters when the cluster structure is known, in this section we simulate data from a bi-clustering model with a TOM structure and estimate it under different scenarios. We simulate data and estimate the model 50 times for each scenario.

In particular, we use a TOM model with three row and two column clusters ($R = 3$ and $C = 2$) and the following parameters for the linear predictor: $\alpha = (0, -3, 3)$, $\gamma = (0, -0.2, 0.5)$, $\beta = (0, 2)$, $\delta = (0, 0.5)$. The model also has the same number of rows and columns on each cluster which implies that $\pi = (1/3, 1/3, 1/3)$ and $\kappa = (0.5, 0.5)$.

We finally set the number of ordinal categories equal to five $(q = 5)$ and the number of columns equal to ten $(p = 10)$. We estimate the TOM under four scenarios with increasing sample size. Scenarios 1-4 present $n = 60$, $n = 120$, $n = 300$ and $n = 1200$, respectively. Table 4.2 shows the true model parameters, the mean and standard deviation (sd) of the simulations for each scenario.

As it can be seen, most means get closer to their true values as the number of rows in the sample (n) increases from Scenario 1 to 4. Further, standard errors also decrease with higher n. For example, in the case of $\beta_2 = 2$ the mean of the estimates goes from 1.85 with a SE of 0.05 in scenario 1 to 2.01 with a SE of 0.01 in Scenario 4. The one exception to the above, is the case of γ_3 which is overestimated even when $n = 1200$. Its true value is 0.5 but the estimated means range from 1.83 to 0.68. Importantly, how-

Table 4.2: Simulation results for TOM bi-clustering with $R = 3, C = 2$ for datasets with $q = 5, p = 10$ and increasing n. Each scenario is based on 50 simulated datasets.

		Scenario 1 $n = 60$		Scenario 2 $n = 120$		Scenario 3 $n = 300$		Scenario 4 $n = 1200$	
Parameter	true	mean	sd	mean	sd	mean	sd	mean	sd
μ_1	-1.95	-2.19	0.08	-2.11	0.08	-1.99	0.04	-1.94	0.01
μ_2	-1.10	-1.24	0.10	-1.14	0.10	-0.98	0.05	-0.91	0.01
μ_3	1.10	0.95	0.14	1.11	0.14	1.32	0.07	1.44	0.01
μ_4	1.95	1.86	0.17	2.06	0.17	2.30	0.08	2.45	0.02
α_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
α_2	-3.00	-2.72	0.69	-2.81	0.43	-2.95	0.21	-2.99	0.08
α_3	3.00	4.33	4.44	3.32	2.01	3.26	0.45	3.16	0.18
γ_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
γ_2	-0.20	-0.27	0.29	-0.29	0.30	-0.20	0.08	-0.20	0.03
γ_3	0.50	1.83	4.20	0.71	0.37	0.62	0.14	0.68	0.19
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_2	2.00	1.85	0.05	1.98	0.05	1.99	0.02	2.01	0.01
δ_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
δ_2	0.50	0.43	0.02	0.44	0.02	0.46	0.01	0.47	0.01
π_1	0.33	0.27	0.01	0.28	0.01	0.31	0.01	0.32	0.00
π_2	0.33	0.32	0.00	0.32	0.01	0.33	0.00	0.33	0.00
π_3	0.33	0.40	0.02	0.40	0.02	0.36	0.01	0.35	0.00
κ_1	0.50	0.49	0.01	0.49	0.01	0.50	0.00	0.50	0.00
κ_2	0.50	0.51	0.01	0.51	0.01	0.50	0.00	0.50	0.00

ever, this does not affect the estimates for the mixture proportions π and κ. The mean of the estimates in each scenario are very close to the true proportions even with small n. As a further illustration, Figure 4.2 shows the estimated parameters for α, γ, β, and δ in Scenarios 1 and 4. They show that the estimated parameters are close to the true values and how they get closer as the number of rows n in the sample increases from 60 to 1200.

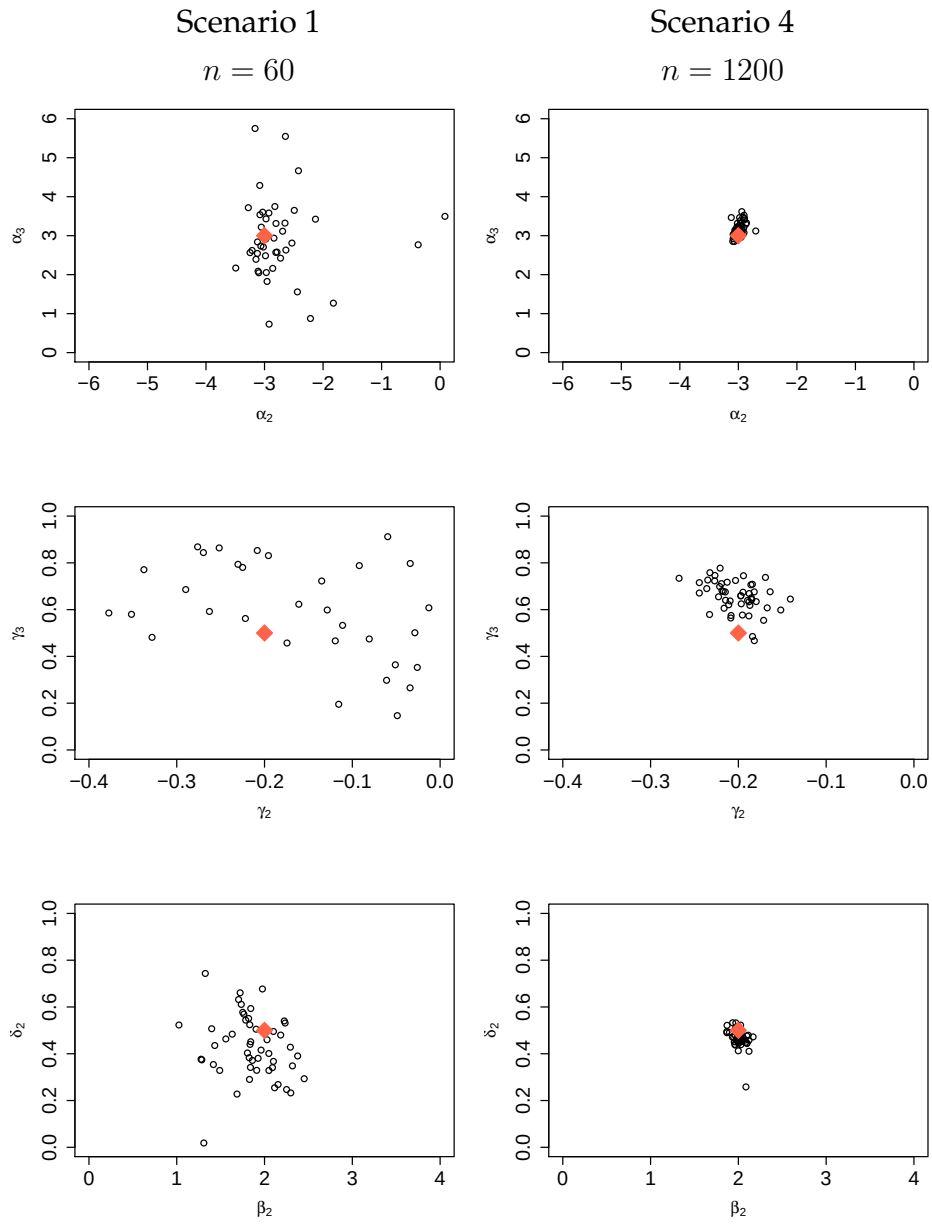


Figure 4.2: Simulation results for the bi-clustering model. True values for α, γ, β and δ are the diamonds in the middle of the plots. Each scenario is based on 50 simulated datasets.

4.4 Case Study: Comparing POM and the TOM using HILDA data

To illustrate the TOM and the POM, we apply them to data from the Household, Income and Labour Dynamics in Australia (HILDA). We use 2001-2011 self-reported health status (SRHS) from this survey. SRHS is an ordinal variable with 5 categories: Poor, Fair, Good, Very Good and Excellent. More details about this dataset can be found in Chapter 2.

As seen in that chapter, SRHS is highly correlated across time. Table 4.3 presents the 2001-2011 transitions between ordinal categories for all individuals. Diagonal proportions are very high, about 40%, and the same is true for the cells close to the diagonal. In words, even after 11 years individuals are very likely to report the same health status or the one next to their starting status.

Table 4.3: SRHS transition matrix 2001-2011

		2011					
		Poor	Fair	Good	Very good	Excellent	Total
2001	Poor	0.42	0.40	0.14	0.04	0.00	1.00
	Fair	0.13	0.44	0.34	0.07	0.01	1.00
	Good	0.02	0.21	0.54	0.20	0.02	1.00
	Very good	0.01	0.09	0.38	0.46	0.07	1.00
	Excellent	0.01	0.04	0.21	0.47	0.27	1.00

In order to illustrate the TOM with clustering, we use a random sample of 136 individuals with complete data over 2001-2011. Therefore, we estimate the models in this chapter using dataset with $n = 136$ rows, $p = 11$ columns and $q = 5$ ordinal levels. In addition to that, we also fit POM with clustering from Chapter 3 and compare the POM and TOM formulations using the Frequentist information criteria introduced in the previous chapter, namely the AIC, BIC and ICL-BIC.

Table 4.4 shows the results. For each fitted model, we present the linear predictor, the number of row (Rows) and columns (Cols) clusters, the total number of parameters (Params), and the AIC, BIC and ICL-BIC.In terms of the latter information criterion, the model with the best fit is the TOM with five row clusters with an ICL-BIC of 2922. It has a total of 16 parameters ($\mu_k, \alpha_r, \gamma_r$ and π_r where $k = 1, \ldots, 4$ and $r = 1, \ldots, 5$). It is closely follow by the POM with five row clusters (ICL-BIC=2927). Note that more parsimonious models are preferred, e.g. row-cluster and not bi-cluster models. On the other hand, the AIC selects the least parsimonious model, a row and column fixed effects model with 149 parameters (AIC=2424) and the BIC selects the TOM with six row-clusters (BIC=2875). The overestimation of the number mixture of components when using the AIC and BIC found by McLachlan & Peel (2000) and Cubaynes et al. (2012) seems to be also shown here.

In sum, results for this random subsample of the 2001-2011 SRHS suggest that the individuals could be grouped into five clusters. What do these estimated five row-clusters look like? The original subsample and the resulting clusters are visualized using heatmaps and are shown in the heatmaps in Figure 4.3. Individuals and occasions are shown in rows and columns and cell colors represent ordinal categories. The five row-clusters comprise: two where SRHS remains stable, two where it slightly improves (each with different starting category) and one where it slightly worsens.

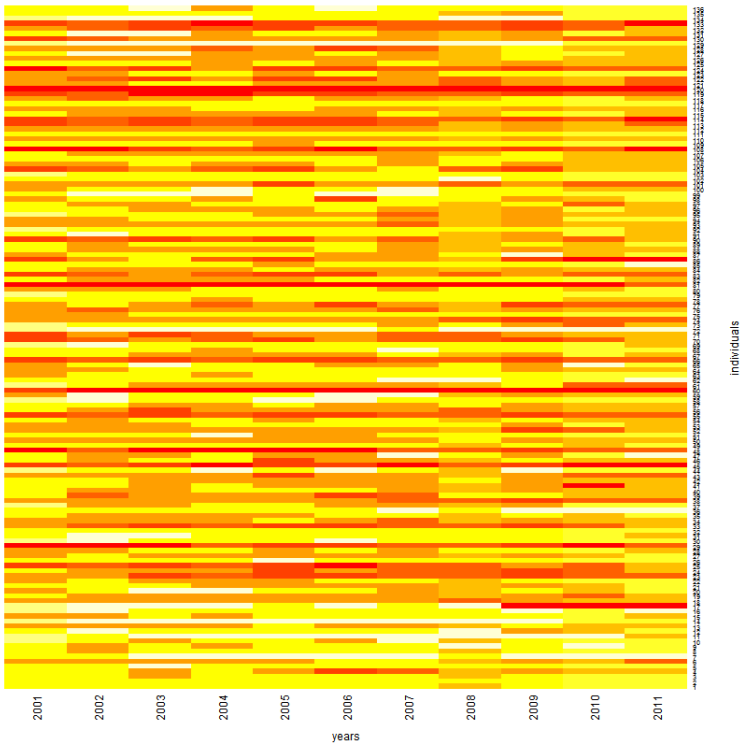
It is important to reiterate that the subsample used in this chapter is only for demonstration purposes and comparison between the proposed versions of the POM and TOM. A more robust analysis, will have to use all the available observations for this dataset and pose further assumptions for the missing data, ie missing at random (MAR) or provide a model the missing data mechanism if the drop-out is non-ignorable (Little & Rubin 2002).

In the chapters to come, Chapters 5, 6, and 7, we will extend some of the models presented so far to explicitly incorporate the correlation over time.

Table 4.4: Model comparison in the case study: SRHS from HILDA

Model	Linear Predictor	Rows	Cols	Param	AIC	BIC	ICL-BIC
Null	μ_k	1	1	4	4089	4139	-
Row effects	$\mu_k - \alpha_i$	136	1	139	2446	4200	-
Col effects	$\mu_k - \beta_j$	1	11	14	4094	4271	-
Row and Col effects	$\mu_k - \alpha_i - \beta_j$	136	11	149	**2424**	4305	-
POM row clustering	$\mu_k - \alpha_r$	2	1	6	3313	3345	3355
		3	1	8	3029	3071	3095
		4	1	10	2920	2973	2999
		5	1	12	2829	2893	2927
		6	1	14	2809	2883	2947
POM row clustering + col effects	$\mu_k - \alpha_r - \beta_j$	2	11	16	3310	3395	3405
		3	11	18	3018	3113	3137
		4	11	20	2904	3010	3038
		5	11	22	2814	2931	2964
		6	11	24	2792	2919	2981
POM col clustering	$\mu_k - \beta_c$	1	2	6	4092	4124	4134
POM bi-clustering	$\mu_k - \alpha_r - \beta_c$	2	2	8	3313	3355	3376
		3	2	10	3024	3077	3111
		4	2	12	2914	2978	3015
		5	2	14	2827	2901	2951
		6	2	16	2803	2888	2967
TOM row clustering	$\mu_k - \alpha_r - \gamma_r t_k$	2	1	7	3313	3351	3360
		3	1	10	3025	3078	3098
		4	1	13	2912	2981	3006
		5	1	16	2808	2893	**2922**
		6	1	19	2774	**2875**	2925
TOM row clustering + col effects	$\mu_k - \alpha_r - \gamma_r t_k - \beta_j$	2	11	17	3308	3398	3407
		3	11	20	3011	3117	3136
		4	11	23	2895	3018	3041
		5	11	26	2791	2929	2958
		6	11	29	2759	2913	2953
TOM col clustering	$\mu_k - \beta_c - \delta_c t_k$	1	2	7	4095	4132	4132
TOM bi-clustering	$\mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c)t_k$	2	2	10	4033	4087	4078
		3	2	13	3882	3951	3938
		4	2	16	3003	3088	3066
		5	2	19	5541	5642	5634
		6	2	22	5776	5893	5885

Unordered data



TOM row-clustered data (5 groups)

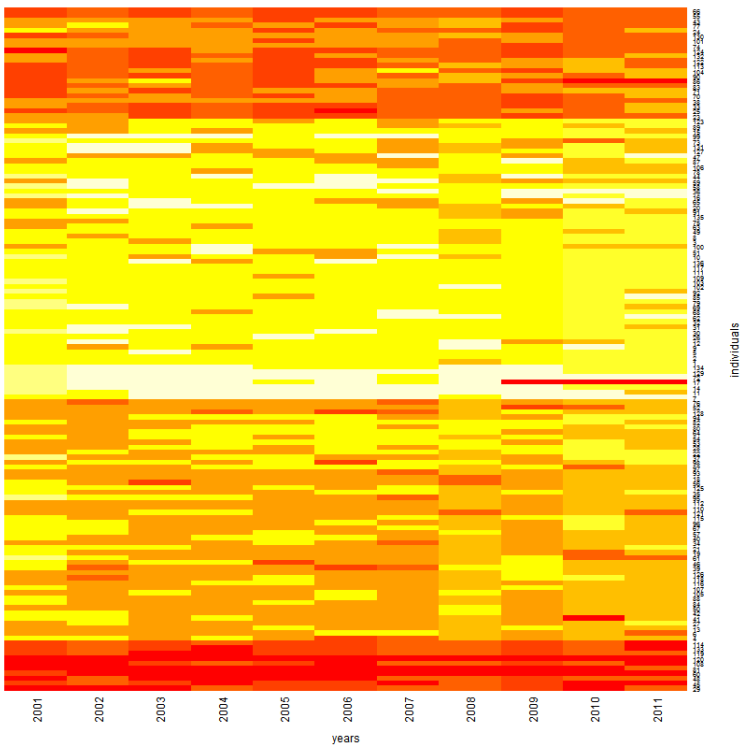


Figure 4.3: Heatmaps for SRHS in HILDA

These models are fitted using a Bayesian approach to take advantage of the flexibility of MCMC methods to estimate models with complex correlation structures. Next, we present a model with latent random effects in Chapter 5.

Part II

Bayesian Estimation

Chapter 5

Parameter dependent models

5.1 Latent random effects models: random walk by cluster

This chapter presents models where the correlation between observations is explicitly modelled through the parameters, an approach that we denominate parameter dependent models in the literature review. This approach introduces the repeated measures correlation by conditioning the response on latent random effects, that is a finite mixture of random effects models (Vermunt et al. 1999, Vermunt & Van Dijk 2001, Bartolucci & Farcomeni 2009, Bartolucci et al. 2014). In particular, we augment the linear predictor of the POM with occasion and cluster specific random effects that follow a random walk with cluster specific variance . Following the same notation as previous chapter, instead of being independent over occasions, now β_{rj} is assumed to arise from $\beta_{rj} \sim N(\beta_{rj-1}, \sigma_r^2)$. The linear predictor for this model, which retains the row-clustering with columns and interaction, is

$$\text{Logit}[P(y_{ij} \leq k | i \in r)] = \mu_k - \alpha_r - \beta_{rj} \tag{5.1}$$

$$
\begin{aligned}
i = 1, \ldots, n;\ j = 1, \ldots, p;\ r = 1, \ldots, R;\ k = 1, \ldots, q \\
\mu_{k+1} \leq \mu_k \text{ for } k = 0, \ldots, q;\ \mu_0 = -\infty;\ \mu_1 = 0;\ \mu_q = \infty \\
\beta_{r1} = 0 \text{ for all } r
\end{aligned}
$$

Notice that following the Bayesian literature for ordinal data (Albert & Chib 1995, Johnson & Albert 1999, Cowles 1996) we are using a slightly different parametrisation in this chapter. In this and subsequent chapters, we set $\mu_1 = 0$ and have no constraint on α_1. By fixing the first cut point, this parametrisation allows better mixing of the MCMC chain.

The model for the repeated ordinal outcomes y_{ij} remains the same as before:

$$
\begin{aligned}
y_{ij} \mid \mu_k, \alpha_r, \beta_{rj}, \pi_r \sim\ & \text{Categorical}_q(\theta_{rjk}) \\
\theta_{rjk} =& \frac{1}{1 + e^{-(\mu_k - \alpha_r - \beta_{rj})}} - \frac{1}{1 + e^{-(\mu_{k-1} - \alpha_r - \beta_{rj})}};\ \sum_{k=1}^{q} \theta_{rjk} = 1 \\
& i = 1, \ldots, n;\ j = 1, \ldots, p;\ r = 1, \ldots, R;\ k = 1, \ldots, q \\
& \mu_{k-1} < \mu_k;\ \mu_1 = 0; \mu_0 = -\infty \text{ and } \mu_q = \infty \\
& \beta_{r1} = 0; \forall r
\end{aligned}
\tag{5.2}
$$

The resulting likelihood is more complex than the corresponding one for the POM or TOM with clustering because it requires integrating out the distribution of the random effects. As mentioned in section 1.3, these integrals are usually solved by numerical methods like the Gauss-Hermite quadrature in the frequentist paradigm. Depending on the number of the quadrature points and the integral's dimension this can be very expensive computationally. Here we take a different approach and carry out estimation using Bayesian methods.

5.2 Bayesian Estimation

In a Bayesian setting, both data and parameters in the model are random variables and thus we need to specify distributions for them. In particu-

lar, in addition to the likelihood which specifies a distribution of the data conditional on the parameters, we need to specify a *prior* distribution for the parameters. By using Bayes theorem, we then obtain the distribution of the parameters given the data, i.e. a *posterior* distribution. Let Y be the data, Ω a set of parameters with prior $P(\Omega)$, the posterior distribution $P(\Omega|Y)$ is given by

$$\begin{aligned} P(\Omega|Y) =& \frac{P(Y|\Omega)P(\Omega)}{P(Y)} = \frac{P(Y|\Omega)P(\Omega)}{\int P(Y|\Omega)P(\Omega)d\Omega} \\ \implies P(\Omega|Y) \propto& P(Y|\Omega)P(\Omega) \end{aligned}$$

where $P(Y|\Omega)$ represents the likelihood and $P(Y)$ the marginal distribution of Y. The latter is also known as marginal likelihood or *evidence* of the model. To complete the specification of our model for y_{ij}, in (5.2), we use the following priors for $\mu, \alpha, \beta, \pi, \sigma^2_\mu, \sigma^2_\alpha$ and σ^2_β :

$$\begin{aligned} \mu_k \mid \sigma^2_\mu \overset{iid}{\sim} \text{Normal}(0, \sigma^2_\mu)\ \text{I}[\mu_k > \mu_{k-1}] \qquad & k = 1, \ldots, (q-1) \\ & \mu_0 = -\infty, \mu_1 = 0, \mu_q = \infty \\ \alpha_r \mid \sigma^2_\alpha \sim \text{Normal}(0, \sigma^2_\alpha) \qquad & r = 1, \ldots, R \\ \beta_{rj} \sim \text{Normal}(\beta_{rj-1}, \sigma^2_r) \qquad & r = 1, \ldots, R;\ j = 1 \ldots p;\ \beta_{r1} = 0,\ \forall r \\ \sigma^2_\mu \sim \text{Inverse Gamma}(a_\mu, b_\mu) & \\ \sigma^2_\alpha \sim \text{Inverse Gamma}(a_\alpha, b_\alpha) & \\ \sigma^2_r \sim \text{Inverse Gamma}(a_\beta, b_\beta) \qquad & r = 2, \ldots, R \\ \pi \sim \text{Dirichlet}(\psi) & \end{aligned} \tag{5.3}$$

with hyperparameters: $\psi = 3/2, a_\mu = a_\alpha = a_\beta = 4$ and $a_\mu = a_\alpha = a_\beta = 1/2$.

In words, we assume that observations y_{ij} come from a hierarchical structure with 3 levels: clusters, individuals and occasions; where only the latter two are observed. The first level of clusters is latent and is where the cluster proportions π_r, the variance of the random effect for each cluster

σ_r^2, and the effect of the cluster in the linear predictor α_r are determined. Next, the level of individuals although observed does not contribute to the linear predictor because we are assuming that all individuals within a cluster are homogeneous, i.e. they have the same probability to have outcome k given that they all are in cluster r. Figure 5.1 shows a graphical representation of the model.

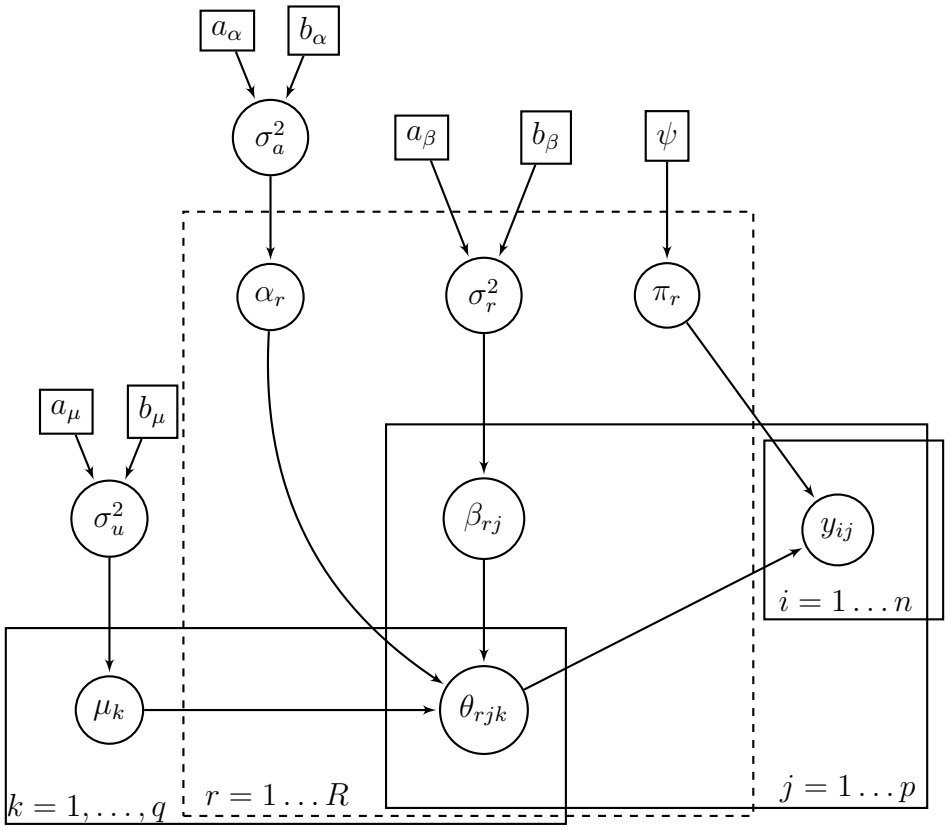


Figure 5.1: Graphical representation of the model

It is important to stress that this formulation assumes that the number of mixture components R is unknown but fixed and thus exogenous to the model. Relaxing this assumption is a natural extension and is explored using a Dirichlet Process Mixture within a Bayesian Non-Parametric approach in Chapter 7.

We implement a Markov-Chain Monte-Carlo (MCMC) sampling scheme that uses the Metropolis-Hastings algorithm (MH), Metropolis et al. (1953), Hastings (1970), to sample from the posterior of the parameters. A Metropolis-Hastings (MH) sampling scheme is necessary because the full conditional distributions for each parameter are non-standard. The model is implemented in R and C++ and is detailed in the next section.

5.3 Construction of the MCMC chain

We now proceed to briefly describe the MH steps, a more detailed treatment could be found in Metropolis et al. (1953) and Chib & Greenberg (1995). The MH algorithm involves the use of an candidate-generating density to sample from any given target. This auxiliary function is often called *proposal* density and is a probability distribution that allows the Markov Chain to move with certain probability from the current state to a new proposed state while maintaining detailed balance, value of the target and proposal evaluated in the current state is equal to the target and proposal evaluated in the new state, also known as reversibility condition. Notice that an initial portion of the chain is discarded as burn-in to allow convergence to the stationary distribution.

Let Ψ be the parameter space of the model of interest, and $\pi(.)$ a target function with proposal $q(.)$. Given a current state $\nu \in \Psi$, we accept a new draw $\nu' \in \Psi$ from the proposal $q(\nu'|\nu)$ with probability r

$$\begin{aligned} r =& \text{min}\left[1, \frac{\pi(\nu')/q(\nu'|\nu)}{\pi(\nu)/q(\nu|\nu')}\right] \\ =& \text{min}\left[1, \frac{P(\nu'|Y)P(\nu')/q(\nu'|\nu)}{P(\nu|Y)P(\nu)/q(\nu|\nu')}\right] \end{aligned}$$

r is also known as Metropolis-Hastings ratio. Notice that in addition to the proposal $q(.)$, the MH ratio actually involves the evaluation of the joint probability of the whole model, likelihood and prior, for both the current

and proposed state. That is, to evaluate $P(Y|\Omega)P(\Omega)$ for $\Omega = \nu, \nu'$ at each iteration of the MCMC chain. As a result, if the calculation of either of these terms is computationally expensive, in particular the likelihood, the target would be slow to simulate. In general, the efficiency of any MH sampler depends on the choice of candidate generating density $q(.)$ and some tuning, trying out different parameters for these proposal, may be needed.

In our case, the posterior distribution of $\Omega = (\mu, \alpha, \beta, \pi, \sigma^2_\mu, \sigma^2_\alpha \sigma^2_\beta)$ is proportional to

$$P(\mu, \alpha, \beta, \pi, \sigma^2_\mu, \sigma^2_\alpha, \sigma^2_\beta | Y) \propto$$
$$P(Y|\mu, \alpha, \beta, \pi) P(\mu|\sigma^2_\mu) P(\sigma^2_\mu) P(\alpha|\sigma^2_\alpha) P(\sigma^2_\alpha) P(\beta|\sigma^2_\beta) P(\sigma^2_\beta) P(\pi)$$

Notice that the first component corresponds to the likelihood and a factorization of the prior that assumes that the likelihood parameters μ, α, β, π are independent. We now proceed to present the proposals and MH ratio for each parameter.

5.3.1 Proposals

Recall the parameter vector $\Omega = (\mu, \alpha, \beta, \pi, \sigma^2_\mu, \sigma^2_\alpha \sigma^2_\beta)$ for the model developed in this chapter (equation 5.2). After choosing initial values for these parameters, for simplicity we use univariate random walk proposals, also known as Metropolis random-walk, and update the current values of each parameter according to the following :

$$
\begin{aligned}
\mu'_k \mid \mu_{k-1}, \mu_k, \mu_{k+1} &\sim U[\,\max(\mu_k - \tau, \mu_{k-1}),\ \min(\mu_k + \tau, \mu_{k+1})] \\
&\qquad k = 2, \ldots, q-1,\ \mu_0 = -\infty, \mu_1 = 0,\ \mu_q = \infty \\
\alpha'_r \mid \alpha_r &\sim \text{Normal}(\alpha_r, \sigma^2_{\alpha p}), \qquad r = 2 \ldots R,\ \alpha_1 = 0 \\
\beta'_j \mid \beta_j &\sim \text{Normal}(\beta_j, \sigma^2_{\beta p}), \qquad j = 2 \ldots p,\ \beta_1 = 0 \\
\text{logit}(w') \mid \text{logit}(w) &\sim \text{Normal}(\text{logit}(w), \sigma^2_{\pi p}) \\
&\qquad w = \pi_{r1}/(\pi_{r1} + \pi_{r2}), \qquad r_1, r_2 \in 1 \ldots R \\
&\qquad \pi'_{r1} = w'(\pi_{r1} + \pi_{r2}) \\
&\qquad \pi'_{r2} = (1 - w')(\pi_{r1} + \pi_{r2}) \\
\log(\sigma'^2_\mu) \mid \log(\sigma^2_\mu) &\sim \text{Normal}(\log(\sigma^2_\mu), \sigma^2_{\sigma\mu p}) \\
\log(\sigma'^2_\alpha) \mid \log(\sigma^2_\alpha) &\sim \text{Normal}(\log(\sigma^2_\alpha), \sigma^2_{\sigma\alpha p}) \\
\log(\sigma'^2_\beta) \mid \log(\sigma^2_\beta) &\sim \text{Normal}(\log(\sigma^2_\beta), \sigma^2_{\sigma\beta p})
\end{aligned}
$$

where the parameters of the proposal densities, also known as *step* size or scale are fixed. In particular, for all the applications in this chapter we use the following step sizes: $\tau = 0.25$, $\sigma^2_{\alpha p} = 0.25$, $\sigma^2_{\beta p} = 1$, $\sigma^2_{\pi p} = 0.1$, $\sigma^2_{\sigma\mu p} = \log(4)$, $\sigma^2_{\sigma\alpha p} = \log(8)$ and $\sigma^2_{\sigma\beta p} = \log(1.5)$.

5.3.2 Acceptance Probabilities (Metropolis-Hastings ratio)

Updates for μ

Choose a μ_k for $k = 2, \ldots, q-1$ at random and sample μ'_k from proposal $q(\mu'_k|\mu_{k-1}, \mu_k, \mu_{k+1})$ and accept with probability

$$
r = \min\left[1, \frac{P(Y|\mu', \alpha, \beta, \pi)P(\mu'|\sigma^2_\mu)}{P(Y|\mu, \alpha, \beta, \pi)P(\mu|\sigma^2_\mu)} \times \frac{\min(\mu_k + \tau, \mu_{k+1}) - \max(\mu_k - \tau, \mu_{k-1})}{\min(\mu'_k + \tau, \mu_{k+1}) - \max(\mu'_k - \tau, \mu_{k-1})}\right]
$$

where $\mu = (\mu_1, \ldots, \mu_k, , \ldots, \mu_{q-1})$, $\mu' = (\mu_1, \ldots, \mu'_k, , \ldots, \mu_{q-1})$ for $k = 1, \ldots, q-1$ and $\mu_1 = 0, \mu_0 = -\infty$, $\mu_q = \infty$.

Updates for α

Choose a $r \in \{1, \ldots, R\}$ at random and sample α_r' from random walk proposal $q(\alpha_r'|\alpha_r)$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, \alpha', \beta, \pi)P(\alpha'|\sigma_\alpha^2)}{P(Y|\mu, \alpha, \beta, \pi)P(\alpha|\sigma_\alpha^2)}\right]$$

where $\alpha = (\alpha_1, \ldots, \alpha_r, , \ldots, \alpha_R)$ and $\alpha' = (\alpha_1, \ldots, \alpha_r', , \ldots, \alpha_R)$.

Updates for β

Choose a r and j from $r = 1, \ldots, R$ and $j = 2, \ldots, p$ at random and sample β_{rj}' from proposal $q(\beta_{rj}'|\beta_{rj})$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, \alpha, \beta', \pi)P(\beta'|\sigma_\beta^2)}{P(Y|\mu, \alpha, \beta, \pi)P(\beta|\sigma_\beta^2)}\right]$$

where $\beta = (0, \ldots, \beta_{rj}, , \ldots, \beta_p)$ and $\beta' = (0, \ldots, \beta_{rj}', \ldots, \beta_p')$.

Updates for σ_μ^2, σ_α^2, σ_β^2

Given σ_μ^2, sample from $\sigma_\mu'^2$ proposal $q(\sigma_\mu'^2|\sigma_\mu^2)$ and accept with probability

$$r = \min\left[1, \frac{P(\beta|\sigma_\mu'^2)P(\sigma_\mu'^2)}{P(\beta|\sigma_\mu^2)P(\sigma_\mu^2)} \times \frac{\sigma_\mu^2}{\sigma_\mu'^2}\right]$$

Similarly for σ_α^2 and σ_β^2.

Updates for π

Given π sample π' from $q(\pi'|\pi)$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, \alpha, \beta, \pi')P(\pi')}{P(Y|\mu, \alpha, \beta, \pi)P(\pi)} \times \frac{w'(1-w')}{w(1-w)}\right]$$

where $\pi = (\pi_1, \ldots, \pi_{r1}, \ldots, \pi_{r2}, \ldots, \pi_R)$, $\pi' = (\pi_1, \ldots, \pi_{r1}', \ldots, \pi_{r2}', \ldots, \pi_R)$, and $w = \pi_{r1}/(\pi_{r1} + \pi_{r2})$, $w' = \pi_{r1}'/(\pi_{r1}' + \pi_{r2}')$

Notice that in the case of α_r and β_{rj} the proposal density $q(.) \sim \text{Normal}(.)$ is symmetric and thus cancels out from the MH ratio. On the other hand, updates for σ^2_μ, σ^2_α, σ^2_β, π involve transformations so that a Jacobian included. Finally, the proposal for μ is not symmetric and thus again it doesn't dropped from the MH ratio.

It is important to mention that when using MH algorithm, the scale of the proposal distributions needs to be tuned in order to get good mixing, that is to maximize the efficiency of the MCMC chain in exploring the target. Smaller steps sizes will have higher acceptances rates but the resulting MCMC chain will slowly explore the target. On the other hand, bigger step sizes might provide better exploration at the expense of lower acceptance rates and thus the chain might be moving slowly as well. Roberts et al. (1997) established that when using the MH algorithm with Gaussian proposals to estimate high-dimensional target distributions, an acceptance rate of 0.234 optimizes the mixing efficiency of the MCMC chain. Step sizes for the proposals in section 6.4 have been tuned so that the acceptance rates are around 20%.

5.3.3 MCMC Convergence

In order to asses the convergence of the MCMC chain, we use the potential scale reduction factor (PSRF) developed by Gelman & Rubin (1992). This statistic is also called Gelman-Rubin convergence diagnostic and unlike other MCMC convergence diagnostics, it is a multiple chain method that uses parallel chains to monitor convergence. In particular, it uses the between-chain and within-chain variances of the marginal posterior of each parameter. Values much higher than one, Gelman & Rubin (1992), indicate a lack of convergence. Given the multimodality in mixture models this diagnostic is particularly helpful in detecting whether or not chains with different starting values have converged to the same mode.

5.4 Model comparison

There are several ways to compare models in a Bayesian framework: (i) using Bayes Factors (Kass & Raftery 1995), (ii) estimating the joint posterior distribution of all competing models using Reversible Jump MCMC (Green 1995, Richardson & Green 1997) and/or other approaches that explore this joint posterior of variable dimension, and (iii) using information criteria. We will use the latter approach here.

Importantly, Frequentist information criteria that use a loss function evaluated at a point estimate are not directly applicable in a Bayesian setting if the posterior distribution of the parameters can not be adequately represented by an unidimensional summary statistic, e.g: mean, median. For example, this is the case for AIC and BIC that compare model (mis)fit by evaluating the log-likelihood at the maximum likelihood estimate. This is especially relevant for mixture models where the likelihood is invariant to the labelling of the individual mixture components and thus the posterior distribution of the parameters is multimodal. This non-identifiability of individual mixture components is a characteristic of mixture models and is known in the literature with the name of the *label switching problem* (McLachlan & Peel 2000, Richardson & Green 1997, Marin et al. 2005).

To compare among competing models we therefore use the Widely Aplicable Information Criterion (WAIC) recently developed by Watanabe (2009). For a model with parameters Ω and data $Y = Y_1, \dots Y_n$, the WAIC is defined as

$$\begin{aligned} \text{WAIC} &= -2 \sum_{i=1}^{n} \log \int P(Y_i|\Omega) P(\Omega|Y) d(\Omega) + 2p \\ &\approx -2 \sum_{i=1}^{n} \log \left[\frac{\sum_{s=1}^{S} P(Y_i|\Omega^s)}{S} \right] + 2p \end{aligned} \tag{5.4}$$

where p is the number of effective parameters

$$\begin{aligned} p &= \sum_{i=1}^{n} \left\{ \log \int P(Y_i|\Omega) p(\Omega|Y) d(\Omega) - \int \log\ P(Y_i|\Omega) p(\Omega|Y) d(\Omega) \right\} \\ &\approx \sum_{i=1}^{n} \left\{ \log \left[\frac{\sum_{s=1}^{S} P(Y_i|\Omega^s)}{S} \right] - \left[\frac{\sum_{s=1}^{S} \log P(Y_i|\Omega^s)}{S} \right] \right\} \end{aligned} \tag{5.5}$$

Watanabe (2009) also provides an alternative approximation for p

$$p_2 = \sum_{s=1}^{S} \text{Var}[\log\ P(Y_i|\Omega)] \tag{5.6}$$

where S is the size of the MCMC chain (number of MCMC draws used for inference). We called these versions WAIC_1 and WAIC_2 depending on whether p_1 or p_2 is used.

Defined this way the WAIC is on the same scale as the AIC and BIC. The contribution of the ith observation to the likelihood $P(Y_i|\Omega)$ is being called *pointwise predictive density* in the literature (Geisser & Eddy 1979, Gelman et al. 2014). We follow this terminology here and further call the WAIC's first component, first term in (5.4), *log predictive density* (LPD). Notice that the WAIC overcomes label switching by integrating out the estimated parameters, posterior $P(\Omega|Y)d(\Omega)$, from the pointwise predictive density $P(Y_i|\Omega)$. In practice, this integral is approximated by MonteCarlo integration using all the MCMC draws $P(Y_i|\Omega^s)$ as shown in the second line of (5.4). A similar procedure is used to approximate the integrals involved in the calculation of p in (5.5).

As a comparison, we also present the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002, 2014) used extensively in Bayesian applications. We separate its two components: Mean Deviance ($\bar{D}$) and number of effective parameters (p_d) so that these could be adequately compared to the WAIC components. However the use of the DIC needs caution when used for singular models, such as in mixtures, hierarchical models and models with a multimodal posterior distribution. In these cases, the num-

ber of effective parameters p_d could be negative and thus the resulting DIC should not be trusted (Celeux et al. 2006, Spiegelhalter et al. 2014).

5.5 Simulations

In this section, we use simulated data to validate both the model estimation and selection procedure. In particular, we simulate one dataset from a three-mixture component with $n = 600, p = 10$ and $q = 5$ and the following parameter values for the mixture model:

$$\begin{aligned}\mu =&(0, 0.85, 3.04, 3.89)\\ \alpha =&(-3, 1, 3)\\ \sigma^2_\beta =&(0.25, 0.50, 0.25)\\ \pi =&(1/3, 1/3, 1/3)\end{aligned}$$

In short, this is a medium size dataset with 600 rows, 10 columns and five ordinal categories; generated from a three latent components with equal proportions. The values for σ^2_β imply that one of the mixture components has a different trend over time (β_{rj}'s are different for $r = 2$). Overall, this model has 41 parameters. We used three chains with over dispersed starting values and ran 7 million iterations for each chain. Discarding the initial 20% draws as burn-in and thinning these chains by 5000, we used for inference 3360 MCMC draws (3 chains of 1120 each).

Figures 5.2 and 5.3 show the marginal posterior and traceplots for all the model parameters. True values are also depicted as vertical/horizontal lines in the graphs. As it can be seen, for all parameters the estimated distribution of the marginal posterior includes the true values and thus the model is able to able to recover all 41 parameters. In terms of mixing, Figure 5.3, shows that MCMC chains exhibit good mixing for all parameters.

A more detailed view of the all 41 estimated parameters can be seen in Table 5.1. This table shows summary statistics for the marginal posteriors of the model parameters (mean, SE and credible intervals), along with the Gelman-Rubin convergence diagnostic (point estimate and upper confidence interval for the PSRF). As seen before, all the credible intervals include the true values of the parameters. In terms of convergence, Gelman-Rubin convergence diagnostics are all close to 1 and well below the threshold value of 1.2, both in terms of the point estimate and upper bound of the confidence interval. The table also shows that these also holds true for the log-likelihood, credible interval includes the likelihood evaluated in the true values and MCMC chains show good mixing.

Table 5.1: Summary statistics for the marginal posterior of the model parameters and Gelman-Rubin convergence diagnostic (PSRF)

				95% Credible Interval		PSRF	
Par	True	Mean	SE	lower	upper	Point.est.	Upper.C.I.
μ_2	0.85	0.86	0.04	0.79	0.94	1.00	1.00
μ_3	3.04	3.09	0.07	2.94	3.23	1.00	1.00
μ_4	3.89	3.91	0.09	3.75	4.08	1.00	1.00
σ^2_μ	–	1.54	0.44	0.82	2.37	1.00	1.00
α_1	-3.00	-2.65	0.27	-3.17	-2.12	1.00	1.01
α_2	1.00	0.88	0.14	0.61	1.15	1.00	1.00
α_3	3.00	3.09	0.14	2.82	3.37	1.01	1.02
σ^2_α	–	1.94	0.63	0.97	3.14	1.00	1.00
β_{12}	-0.08	-0.08	0.30	-0.64	0.55	1.00	1.01
β_{13}	-0.05	-0.33	0.36	-1.01	0.40	1.01	1.02
β_{14}	0.05	-0.32	0.36	-1.06	0.35	1.00	1.01
β_{15}	0.06	-0.56	0.39	-1.29	0.21	1.00	1.01
β_{16}	0.19	-0.48	0.39	-1.26	0.27	1.00	1.01
β_{17}	0.27	-0.34	0.38	-1.05	0.41	1.00	1.01
β_{18}	0.44	0.00	0.34	-0.62	0.73	1.00	1.00
β_{19}	0.54	0.35	0.34	-0.25	1.07	1.00	1.00

Continued on next page...

				95% Credible Interval		PSRF	
Par	True	Mean	SE	lower	upper	Point.est.	Upper.C.I.
β_{110}	0.41	-0.08	0.39	-0.88	0.65	1.00	1.00
β_{22}	-0.71	-0.60	0.17	-0.94	-0.27	1.00	1.00
β_{23}	-0.72	-0.78	0.18	-1.16	-0.45	1.00	1.01
β_{24}	-1.36	-1.09	0.18	-1.45	-0.73	1.00	1.00
β_{25}	-1.42	-1.39	0.19	-1.76	-1.03	1.00	1.00
β_{26}	-1.80	-1.48	0.20	-1.88	-1.12	1.00	1.00
β_{27}	-1.28	-1.12	0.18	-1.47	-0.76	1.00	1.00
β_{28}	-1.15	-1.10	0.18	-1.46	-0.74	1.00	1.01
β_{29}	-0.88	-0.80	0.18	-1.14	-0.45	1.00	1.01
β_{210}	-0.59	-0.55	0.18	-0.91	-0.19	1.00	1.01
β_{32}	0.05	-0.17	0.16	-0.47	0.14	1.00	1.01
β_{33}	-0.31	-0.40	0.17	-0.72	-0.06	1.01	1.03
β_{34}	-0.63	-0.52	0.17	-0.87	-0.20	1.01	1.02
β_{35}	-0.68	-0.63	0.17	-0.95	-0.27	1.00	1.01
β_{36}	-0.48	-0.55	0.17	-0.89	-0.23	1.00	1.01
β_{37}	-0.97	-0.77	0.17	-1.10	-0.44	1.01	1.02
β_{38}	-0.88	-0.92	0.17	-1.23	-0.57	1.01	1.02
β_{39}	-1.03	-1.10	0.18	-1.46	-0.75	1.01	1.02
β_{310}	-0.79	-0.82	0.17	-1.18	-0.50	1.01	1.02
$\sigma^2_{\beta 1}$	0.25	0.44	0.16	0.18	0.75	1.03	1.08
$\sigma^2_{\beta 2}$	0.50	0.42	0.11	0.24	0.64	1.01	1.04
$\sigma^2_{\beta 3}$	0.25	0.33	0.10	0.18	0.53	1.00	1.00
π_1	0.33	0.34	0.02	0.31	0.37	1.00	1.00
π_2	0.33	0.33	0.01	0.30	0.36	1.00	1.00
π_3	0.33	0.33	0.01	0.30	0.36	1.00	1.00
log-like	-6158.29	-6160.99	3.86	-6168.75	-6153.86	1.00	1.00

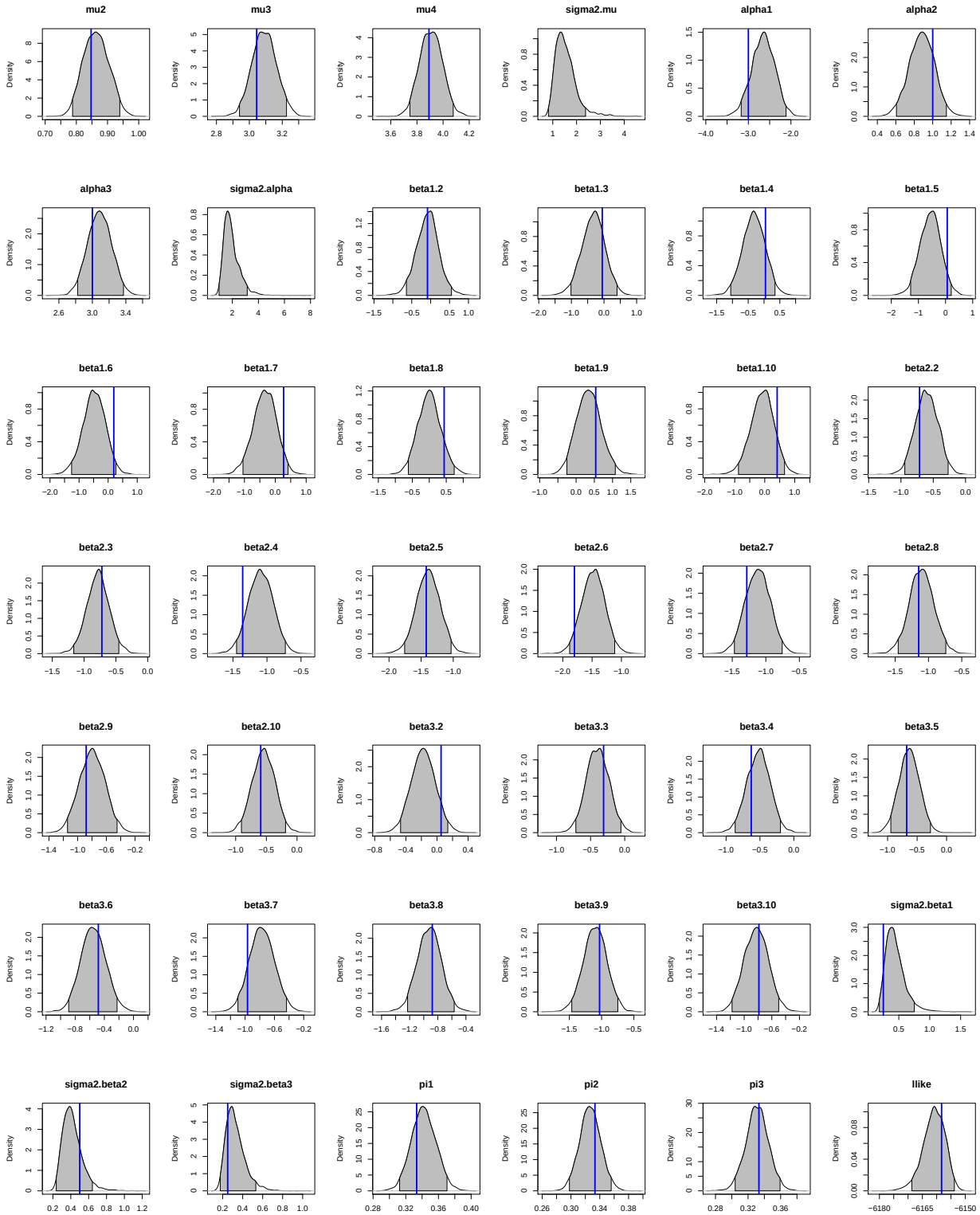


Figure 5.2: Marginal posterior for all models parameters. Vertical line indicates true value

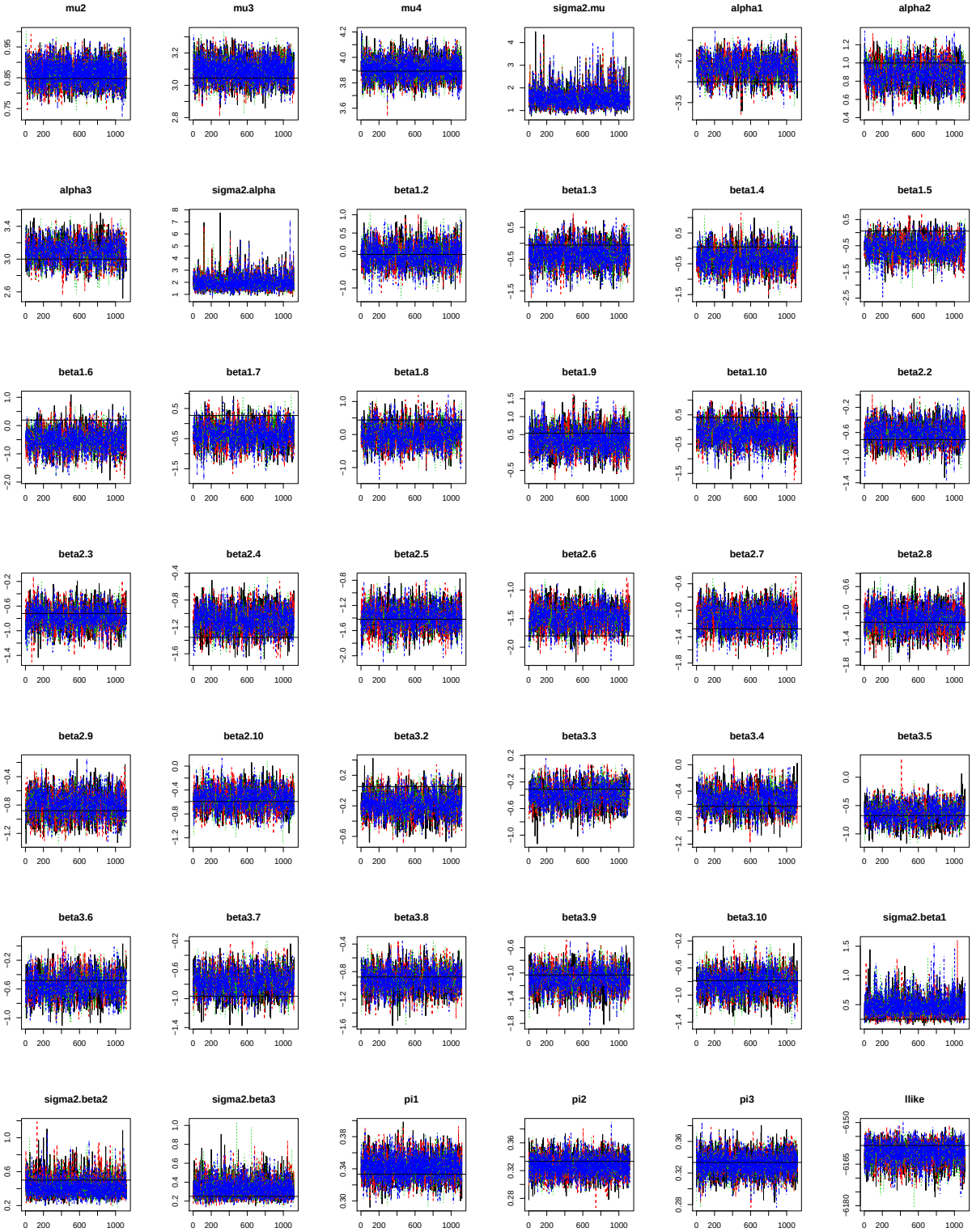


Figure 5.3: Traceplots for all models parameters. Horizontal lines indicate true values. Each MCMC chain is plotted separately

5.5.1 Model comparison

This section compares different models for the simulated dataset using the WAIC (see section 5.4). To do so, we fitted models with a varying number of mixture components from $R = 1$ to $R = 5$ to the simulated dataset, with three mixture components, we used before. Table 5.2 presents the results.

Table 5.2: Bayesian model comparison using WAIC and DIC for simulated data where $\beta_{rj} \sim N(\beta_{rj-1}, \sigma_r^2)$

R	pars	$\overline{D}$	p_{DIC}	DIC	LPD	p_{WAIC1}	WAIC_1	p_{WAIC2}	WAIC_2
1	15	15411	13	15425	15395	16	15427	16	15428
2	29	12916	-945	11971	12891	24	12940	25	12941
3	41	12322	-615	11707	12294	28	12350	28	12350
4	53	12321	-1525	10796	12290	31	12352	30	12351
5	65	12320	-3025	9296	12287	34	12354	32	12352

Here we can see that the model with the lowest WAIC is the model where $R = 3$, as for both versions WAIC_1=WAIC_2=12350 the lowest. The WAIC therefore correctly identifies the true model. On the other hand, the table also shows that the DIC is not a good information criterion in this case. The estimated number of effective parameters (p_{DIC}) is negative whenever $R \geq 2$, that is when mixture models are fitted. This is expected as the DIC should not be use with singular models (Celeux et al. 2006, Spiegelhalter et al. 2014) as discussed in section 5.4.

5.5.2 Estimating a model with misspecified random effects

In order to assess the usefulness of the model proposed in this chapter, we also estimate a model with misspecified distribution of β_{rj}. That is, we use the same simulated dataset but now fit a model where the occasion random effects by cluster are independent. More precisely, we use a prior $\beta_{rj} \sim \text{Normal}(0, \sigma^2_{\beta_r})$ when the data is being generated under

$\beta_{rj} \sim N(\beta_{rj-1}, \sigma_r^2)$. Figure 5.4 compares the estimates for some of the model parameters under both scenarios.

Importantly, the estimates for the cluster effects α and occasion effects by cluster β_{rj} are very similar under both scenarios. This can be seen in the upper panels of Figure 5.4 that show the time trends in the cluster effects by cluster, mean marginal posterior for $\alpha_r + \beta_{rj}$, for both the correct and misspecified priors. Regardless of small differences the time trends are very similar. This is also the case for most model parameters, including μ and π. Complete results for all parameters under the misspecified random effects could be found in Table 5.8 in the Appendix at the end of the chapter.

In contrast to that, the estimates for σ_r^2 differ significantly. The bottom panel in Figure 5.4 shows the marginal posteriors for σ_r^2 under the correct and misspecified random effects. It is evident that these distributions are different, with the estimates of the latter being much larger. In words, the variances of the occasion effects by cluster are bigger when they are assumed to be independent. Moreover, the marginal posteriors under the misspecified model do not include the true values for σ_r^2.

Would it be possible to detect this misspecification in real datasets? Which of these priors fits the data better? To answer this question, we use the WAIC to compare between them. Table 5.3 presents the WAIC for the misspecified models. For $R = 3$ we can see that the WAIC for the misspecified model is 12360 (WAIC_1=WAIC_2), which is higher than the WAIC for the model under the correct prior 12350 (again WAIC_1=WAIC_2) in Table 5.2 . Moreover, this is also the case for all $R = 1$ to $R = 5$. That is, regardless of the number of mixture components, the WAIC for a model with the correct prior (Table 5.2) is always lower than the WAIC for a model with a misspecified prior (Table 5.3). In sum, for the aforementioned simulated dataset the WAIC allows us not only to identify the number of mixture components but also the correct distribution of β_{rj}.

$\alpha_r + \beta_{rj}$

Correct

Misspecified

σ_r^2

Correct

Misspecified

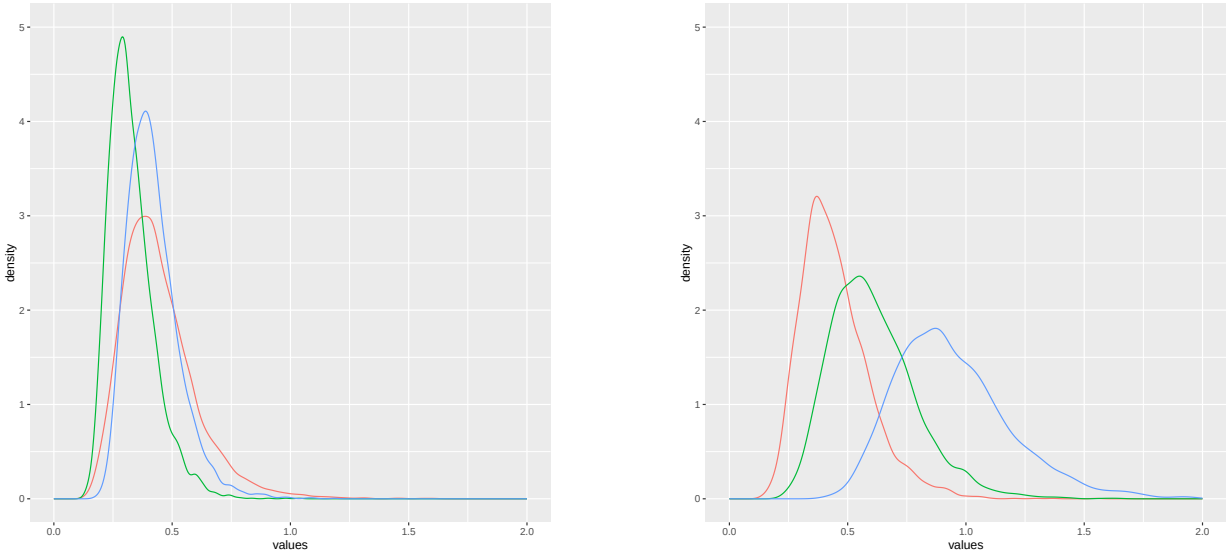


Figure 5.4: Estimated time trends (mean marginal posterior for $\alpha_r + \beta_{rj}$) and marginal posterior distributions for σ_r^2 for simulated data fitted under both correct (left panels) and misspecified (right panels) priors

Table 5.3: Bayesian model comparison using WAIC and DIC for simulated data for model with prior $\beta_{rj} \sim \text{Normal}(0, \sigma^2_{\beta_r})$ when true model has $\beta_{rj} \sim N(\beta_{rj-1}, \sigma^2_r)$

R	pars	$\overline{D}$	p_{DIC}	DIC	LPD	p_{WAIC1}	WAIC_1	p_{WAIC2}	WAIC_2
1	15	15415	15	15430	15398	17	15432	17	15432
2	29	12924	-940	11985	12897	27	12951	27	12952
3	41	12328	-616	11712	12297	32	12360	32	12360
4	53	12327	-3181	9146	12292	35	12362	34	12361
5	65	12326	-3838	8488	12289	37	12364	36	12362

5.6 Case Study: 2009-2013 life satisfaction in New Zealand

We now estimate the model using life satisfaction in New Zealand over 2009-2013 from the New Zealand Attitudes and Values Survey (NZAVS). Life satisfaction is an ordinal variable with seven levels: 1 (Strongly disagree) to 7 (Strongly Agree) and it is been recorded for 2564 subjects. This dataset thus have $n = 2564$, $p = 5$ and $q = 7$. More details of this dataset could be found in Chapter 3.

We also already analysed this dataset using Frequentist methods (EM algorithm) in Chapter 3 using a model where the occasion effects β_j where independent and did not vary by cluster. In Bayesian terms, such a model can be fitted using $\beta_j \sim \text{Normal}(0, \sigma^2_\beta), j = 1 \dots p$ as prior for the occasions.

5.6.1 Model comparison

We fit the model with varying number of mixture components from $R = 1$ to $R = 6$. In this case, we used five chains with over dispersed starting values and ran the MCMC chain for 7 million iterations. Discarding the initial 40% draws as burn-in and thinning these chains by 2500, we used for inference 8400 MCMC draws (5 chains of 1680 each). Table 5.4 presents

the results.

Firstly, notice that due to label switching the DIC can not be trusted in this case. For models where $R > 2$, the number of effective parameters (p_{DIC}) is negative and therefore the value of the DIC is underestimated. For the NZAVS data the WAIC decreases monotonically, both for WAIC1 and WAIC2, for all values of R. This means that the model with the lowest WAIC is the model with six components. This model is not parsimonious at all as it has 70 parameters. However, we do not choose the model with $R = 6$ because as an Bayesian information criteria equivalent to the AIC, the WAIC will tend to overestimate the number of mixture components. In fact, as seen in Table 3.9 of Chapter 3, for the NZAVS dataset the AIC chooses a model with individual row and column effects with 2573 parameters, that is the least parsimonious model of all considered.

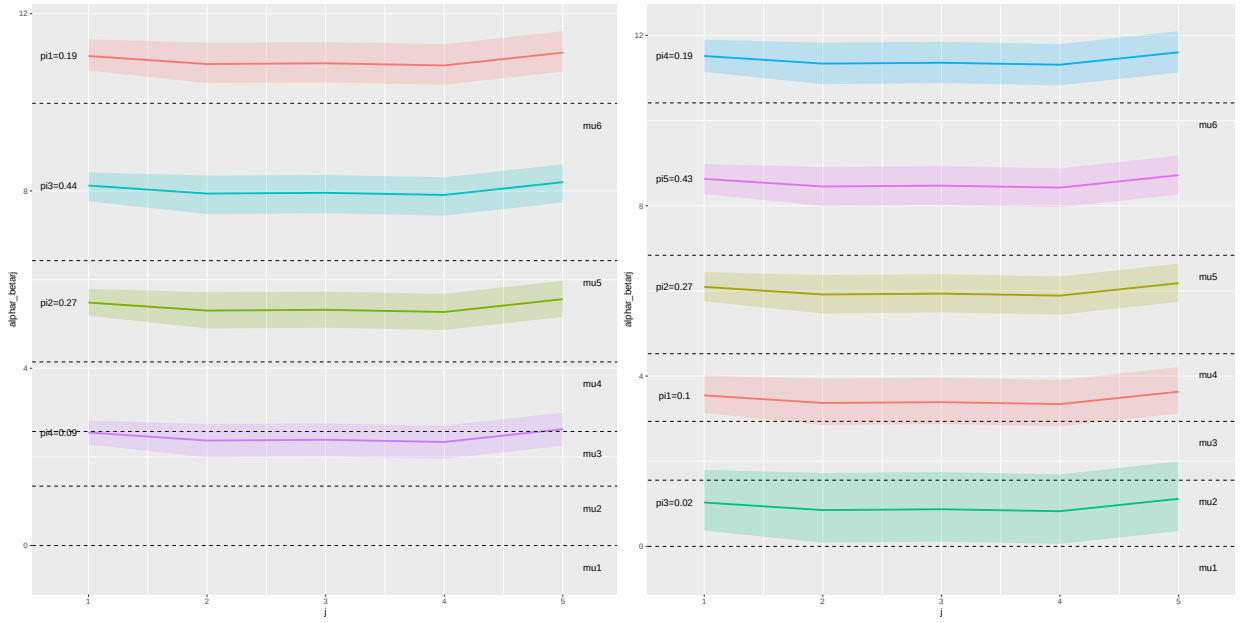
Given this tendency of the WAIC to select models with a large number of mixture components, just like the AIC, we decide to present the results for a more parsimonious model as an illustration of the model in this dataset. We then present the results for the model with R=4, which has the lowest ICL-BIC, Frequentist information criteria, amongst similar models considered then (Table 3.9, Chapter 3).

Table 5.4: Bayesian model comparison using WAIC and DIC for latent random effects models for the NZAVS data

Model	R	pars	$\overline{D}$	p_{DIC}	DIC	LPD	p_{WAIC1}	WAIC_1	p_{WAIC2}	WAIC_2
	1	11	37536	19	37555	37505	31	37567	31	37568
	2	16	33144	18	33162	33111	33	33176	33	33176
	3	18	31436	-3056	28380	31405	31	31467	31	31467
$\beta_j \sim \text{Normal}(0, \sigma^2_\beta)$	4	20	30731	-3650	27081	30702	29	30760	29	30760
	5	22	30612	-1807	28806	30575	37	30650	38	30651
	6	24	30560	-5376	25184	30519	41	30601	41	30602
	1	11	37536	19	37555	37505	31	37567	31	37568
	2	21	33138	22	33160	33104	34	33172	35	33173
	3	28	31426	-3058	28368	31390	36	31463	37	31463
$\beta_{rj} \sim \text{Normal}(0, \sigma^2_{\beta_r})$	4	35	30721	-2834	27887	30682	39	30760	39	30760
	5	42	30601	-2518	28083	30546	55	30656	55	30656
	6	49	30544	-12972	17573	30481	64	30608	64	30610
	1	11	37536	19	37555	37505	31	37567	31	37567
	2	16	33144	18	33162	33111	33	33177	33	33177
	3	18	31436	-3057	28379	31405	31	31467	31	31467
$\beta_j \sim \text{Normal}(\beta_{j-1}, \sigma^2_\beta)$	4	20	30731	-3651	27080	30702	29	30760	29	30760
	5	22	30612	-1325	29287	30575	37	30649	38	30650
	6	24	30559	-4414	26145	30510	49	30608	50	30610
	1	11	37536	19	37555	37505	31	37567	31	37567
	2	21	33138	22	33160	33104	34	33172	34	33173
	3	28	31426	-3059	28367	31389	37	31464	37	31464
$\beta_{rj} \sim \text{Normal}(\beta_{rj-1}, \sigma^2_{\beta_r})$	4	35	30720	-4286	26435	30682	39	30759	39	30760
	5	42	30599	-2529	28070	30545	54	30653	54	30653
	6	49	30542	-7025	23517	30477	65	30607	66	30609

$$\beta_j \sim \text{Normal}(0, \sigma^2_\beta)$$

$R = 4$ $R = 5$



$$\beta_{rj} \sim \text{Normal}(\beta_{rj-1}, \sigma^2_{\beta_r})$$

$R = 4$ $R = 5$

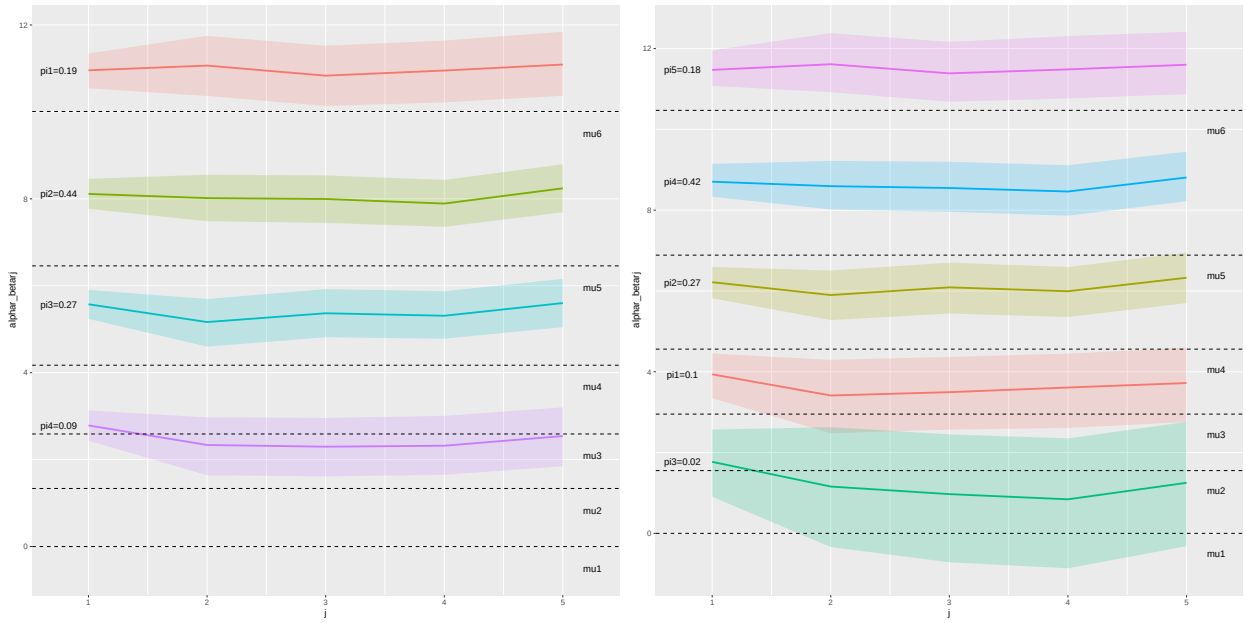


Figure 5.5: Estimated time trends for the models with the lowest WAIC $R = 4$ and $R = 5$

5.6.2 Parameter estimates

Table 5.5 presents the results for the NZAVS data for a model with four mixture components. For this model, participants in the 2nd cluster have the highest levels of life satisfaction ($\alpha_2 = 10.92$) and represent about a fifth of the sample ($\pi_2 = 0.19$) whereas those in cluster one have the lowest levels ($\alpha_1 = 2.8$) and are also the smallest cluster ($\pi_1 = 0.09$). On the other hand, cluster three is the biggest and has high levels of life satisfaction ($\alpha_3 = 8.07, \pi_3 = 0.44$). A noteworthy feature of this model is that the estimated variances of the random effects by cluster, $\sigma^2_{\beta_r} = (0.46, 0.42, 0.41, 0.43)$ are very similar to each other, pointing out that a more parsimonious alternative might be preferred.

In addition to that, by plotting cluster effects in the linear predictor overtime ($\alpha_r + \beta_{rj}$) we can see that they seem to be parallel, confirming that there is no need to include interactions β_{rj} as suggested by the Frequentist model comparison in Chapter 3 (Table 3.9) where models with columns effect only are preferred over model with column effect and interactions.

5.6.3 Classification results

We finally present in Figures 5.6 and 5.7 the classification results for the model with four components. The former presents the probability that each respondent belongs to the same cluster over all MCMC chains (Co-clustering Probabilities) for both the original and clustered data. We can see that when ordering the respondents by cluster, we are able to visualise higher co-clustering probabilities within cluster as well as their relative size, i.e. estimated cluster proportions $\pi = (0.09, 0.19, 0.44, 0.27)$.

Figure 5.7 on the other hand, shows the distribution of life satisfaction over 2009-2013 for each of the estimated latent groups. As remarked before, the 2nd cluster has the highest levels of life satisfaction and the 1st one the lowest. It is also noticeable that there is not much variation over time within each cluster, confirming the results we already obtained for

Table 5.5: Summary statistics for the marginal posteriors and Gelman-Rubin convergence diagnostic (PSRF) for the NZAVS dataset with $R = 4$.

			95% Credible Interval		PSRF	
Par	Mean	SE	lower	upper	Point.est.	Upper.C.I.
μ_2	1.32	0.10	1.13	1.51	1.00	1.00
μ_3	2.56	0.11	2.36	2.79	1.00	1.01
μ_4	4.14	0.12	3.92	4.40	1.00	1.01
μ_5	6.43	0.13	6.18	6.68	1.00	1.01
μ_6	9.97	0.14	9.70	10.23	1.00	1.00
σ^2_μ	3.14	0.71	1.92	4.49	1.00	1.00
α_1	2.80	0.16	2.50	3.11	1.01	1.01
α_2	10.92	0.17	10.57	11.25	1.00	1.00
α_3	8.07	0.15	7.76	8.35	1.00	1.01
α_4	5.54	0.14	5.28	5.83	1.00	1.00
σ^2_α	5.98	1.75	3.43	9.56	1.00	1.00
β_{12}	-0.44	0.16	-0.76	-0.14	1.00	1.01
β_{13}	-0.54	0.16	-0.85	-0.22	1.00	1.01
β_{14}	-0.49	0.17	-0.81	-0.17	1.00	1.01
β_{15}	-0.28	0.17	-0.62	0.03	1.00	1.01
β_{22}	0.13	0.15	-0.15	0.43	1.00	1.01
β_{23}	-0.10	0.15	-0.39	0.19	1.00	1.01
β_{24}	-0.00	0.16	-0.32	0.29	1.00	1.01
β_{25}	0.14	0.16	-0.17	0.47	1.00	1.01
β_{32}	-0.08	0.10	-0.26	0.12	1.00	1.00
β_{33}	-0.10	0.10	-0.30	0.10	1.00	1.01
β_{34}	-0.21	0.10	-0.41	-0.01	1.00	1.01
β_{35}	0.15	0.10	-0.05	0.35	1.00	1.01
β_{42}	-0.40	0.11	-0.61	-0.19	1.00	1.00
β_{43}	-0.21	0.11	-0.42	0.00	1.00	1.01
β_{44}	-0.25	0.11	-0.47	-0.04	1.00	1.00
β_{45}	0.03	0.11	-0.20	0.25	1.00	1.00
$\sigma^2_{\beta 1}$	0.46	0.18	0.19	0.83	1.01	1.04
$\sigma^2_{\beta 2}$	0.42	0.17	0.16	0.75	1.00	1.00
$\sigma^2_{\beta 3}$	0.41	0.17	0.16	0.73	1.02	1.05
$\sigma^2_{\beta 4}$	0.43	0.17	0.18	0.76	1.03	1.07
π_1	0.09	0.01	0.08	0.10	1.00	1.00
π_2	0.19	0.01	0.17	0.21	1.00	1.00
π_3	0.44	0.01	0.42	0.47	1.00	1.00
π_4	0.27	0.01	0.25	0.29	1.00	1.00
log-like	-15628.98	3.72	-15636.43	-15622.10	1.00	1.00
log-post	-15670.82	4.18	-15678.97	-15662.88	1.00	1.00

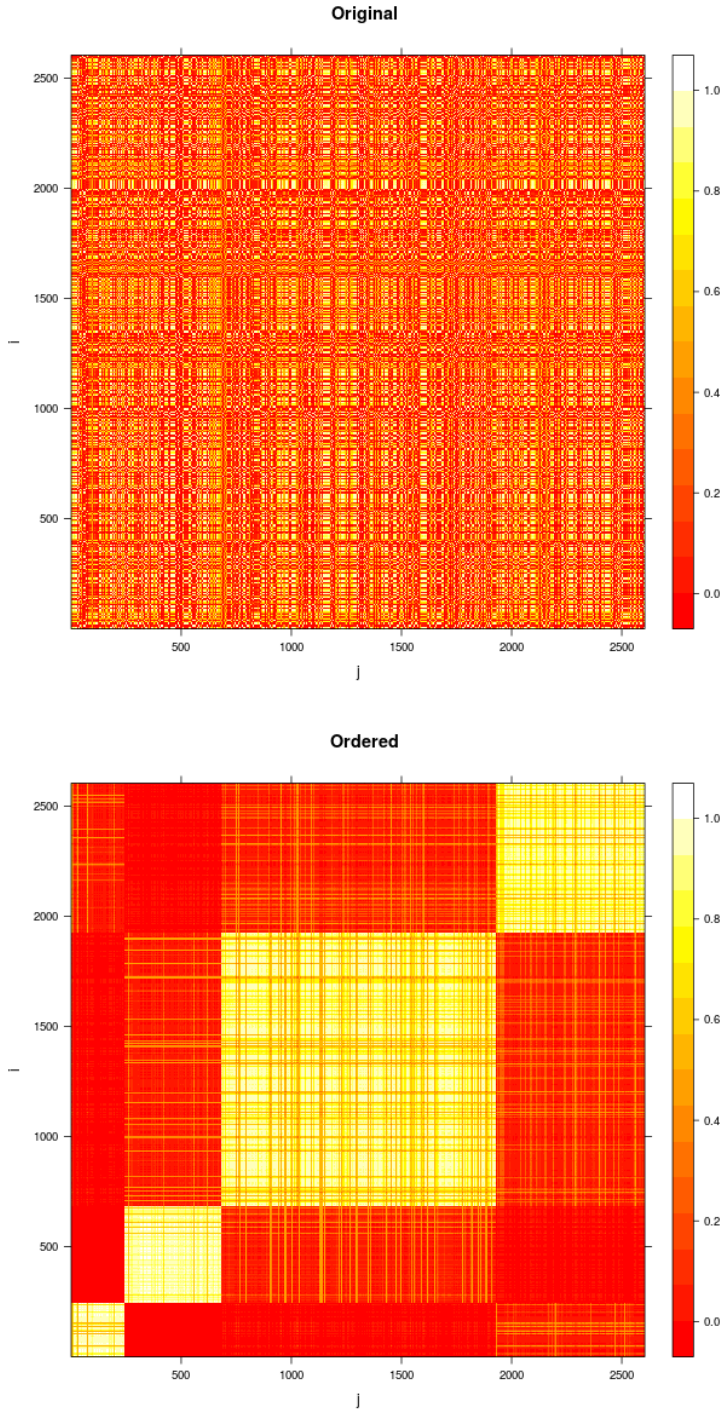


Figure 5.6: Co-clustering Probabilities (Mean Posterior) in the original and ordered data for the NZAVS dataset

the NZAVS data in this and earlier chapters.

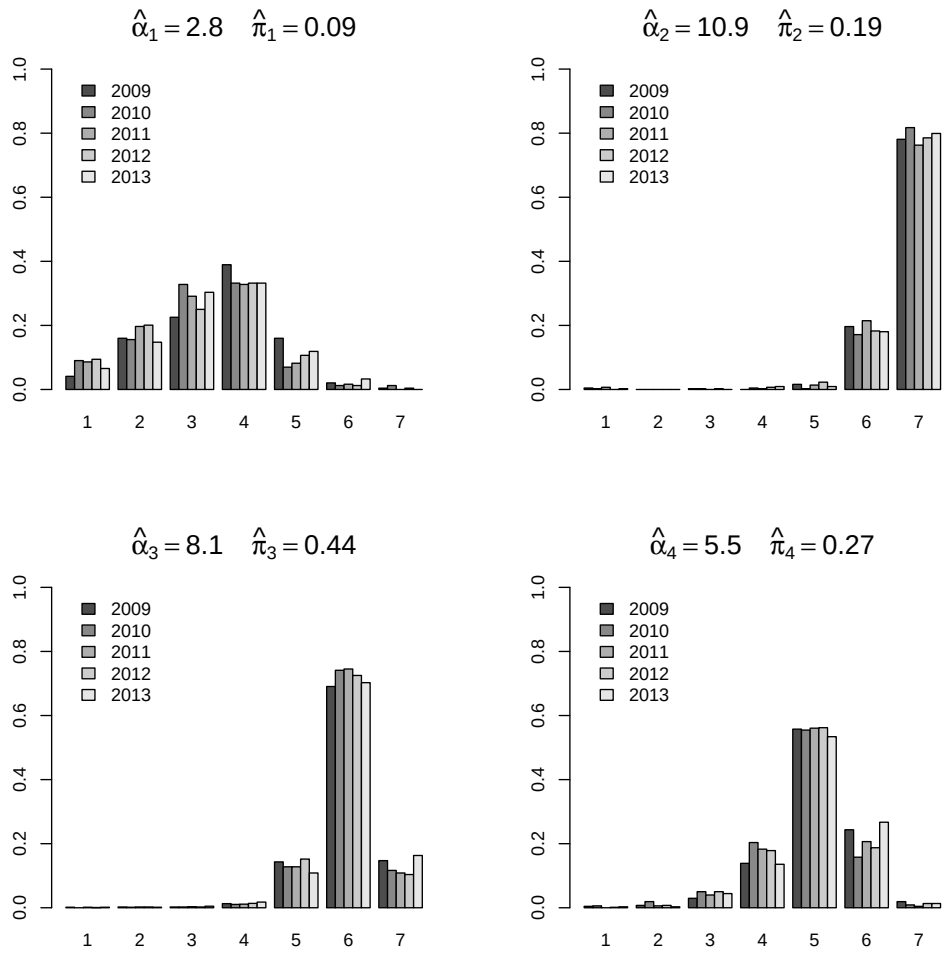


Figure 5.7: Life Satisfaction distribution by estimated cluster

5.7 Case Study: Variant strains in infant gut bacteria

In this section, we identify variant strains of Bacteroides Faecis (B. faecis) using data from baby stool samples. Variant strains are dominant genotypes, similar alleles configurations in a given genome, and are identified comparing sequenced data from the sample to a reference genome. Although the original shotgun metagenomic data are counts, number of reads that are equal and differ to a reference genome we use the following ordinal version

- "reference": includes SNV sites where all reads are the same as the reference (fixed to reference);
- "segregating" more than 5 reads are equal to reference and more than 5 reads are equal to a different one (segregating site);
- "non-reference": all reads for the cell are the same allele which is different to the reference one (fixed to non-reference).

The data thus comprises an ordinal response with three levels measured for 1992 sites over 25 occasions in one infant. Figure 5.8 shows a graphical representation of the this data. More details of this dataset can be found in Chapter 2.

5.7.1 Model comparison

For the infant gut dataset, we estimate the model for $\beta_{rj} \sim \text{Normal}(0, \sigma^2_{\beta_r})$ and $\beta_{rj} \sim \text{Normal}(\beta_{rj-1}, \sigma^2_{\beta_r})$ with a varying number of mixture components from $R = 1$ to $R = 4$. In this case, we used five chains with over dispersed starting values and ran the MCMC chain for 3 million iterations. Discarding the initial 30% draws as burn-in and thinning these chains by 1250, we used for inference a chain with 8400 MCMC draws (5 chains of

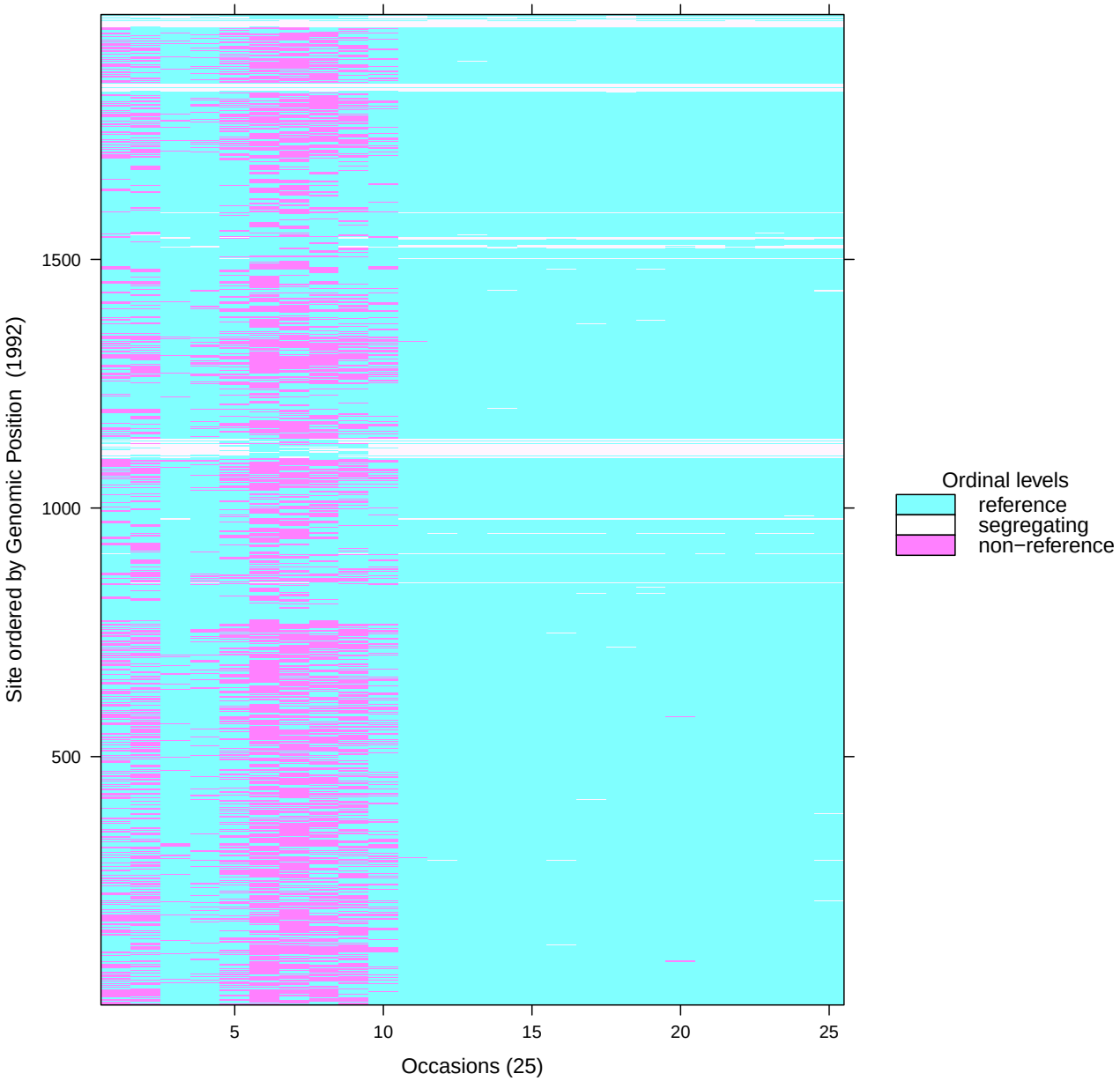


Figure 5.8: Heatmap for the infant gut bacteria dataset

1680 each). In contrast to the previous case study, we relabelled the MCMC chains using the algorithm proposed by Stephens (2000) for all the models so that the DIC could also be used. In addition to that, given the patterns display in the heatmap we only estimate models with occasions and cluster interactions (β_{rj}). Table 5.6 presents both versions of the WAIC and DIC as well as their components.

We can see that all these Bayesian information criteria decrease monotonically with R, suggesting the least parsimonious models (111 parameters). Notice also that all information criteria suggest a slight preference for models where β_{rj} are correlated within cluster.

Table 5.6: Bayesian model comparison using WAIC and DIC for latent random effects models for the infant gut dataset

Distribution for β_{rj}	R	pars	$\overline{D}$	p_{DIC}	DIC	LPD	p_{WAIC1}	WAIC_1	p_{WAIC2}	WAIC_2
Normal$(0, \sigma^2_{\beta_r})$	1	27	41847	27	41874	41799	48	41896	49	41896
	2	57	35119	43	35162	35053	66	35184	69	35191
	3	84	33531	61	33592	33445	86	33617	91	33627
	4	111	33172	75	33246	33068	103	33275	111	33289
Normal$(\beta_{rj-1}, \sigma^2_{\beta_r})$	1	27	41847	26	41873	41798	49	41895	49	41896
	2	57	35112	36	35148	35047	64	35176	67	35181
	3	84	33526	52	33578	33442	84	33610	87	33617
	4	111	33174	66	33239	33073	101	33275	106	33285

Although a better fit for the data, would a model with four variant strains be more meaningful? Figure 5.9 compares it with a model with three strains. As it can be seen, the former provides a better fit but their estimates have higher uncertainty. Similarly to the NZAVS case study, a frequentist estimation of similar models led to the same conclusion (model selected by BIC and ICL was $R = 3$ whereas AIC selected a models with higher R). In light of this, we choose a model with only three variant strains.

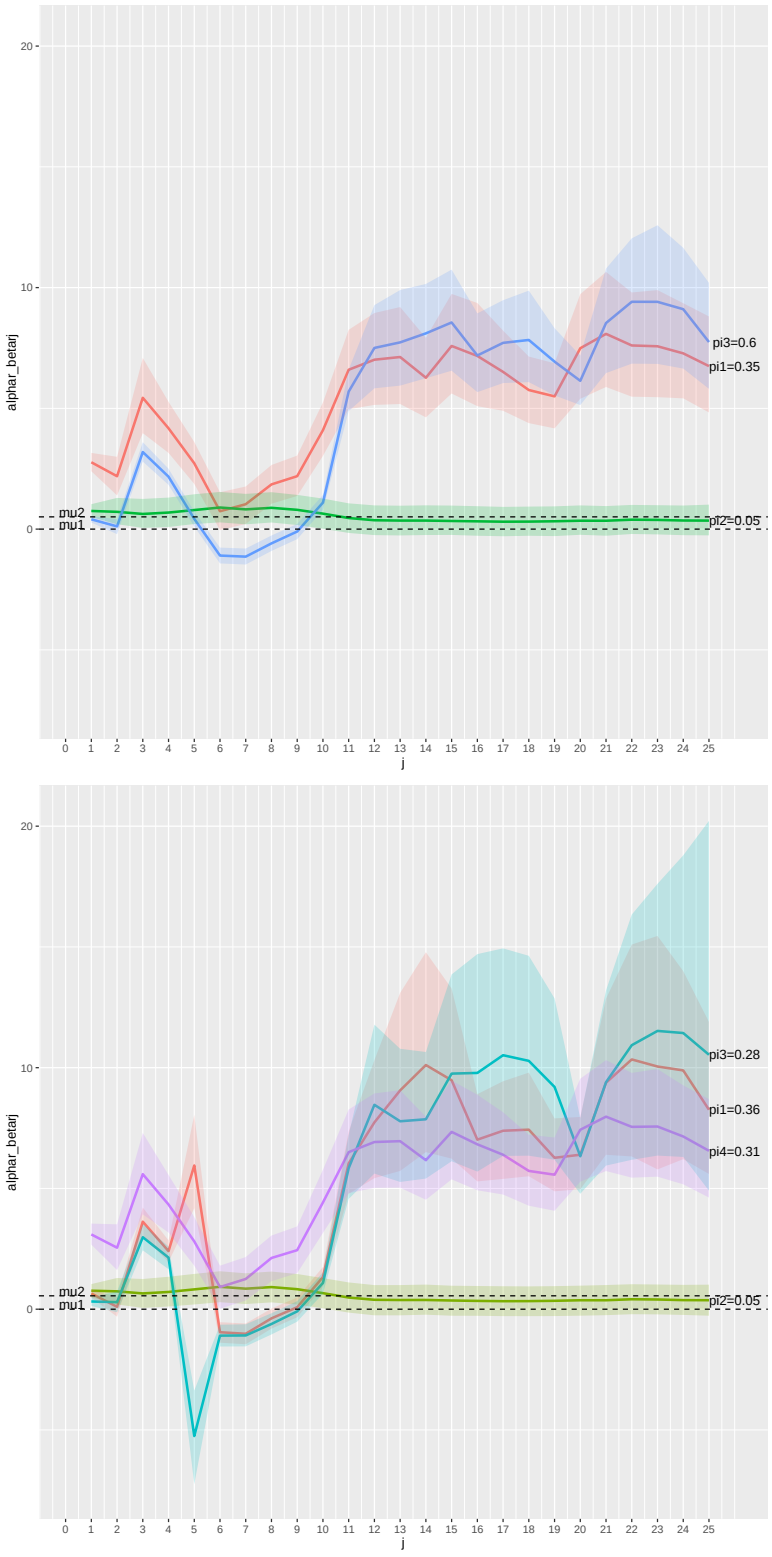


Figure 5.9: Estimated time trends for $\beta_{rj} \sim \text{Normal}(\beta_{rj-1}, \sigma^2_{\beta_r})$ models with $R = 3$ and 4 for the infant gut bacteria dataset.

5.7.2 Parameter estimates

Table 5.7 we present the results for the selected model with three strains. For this model, the smallest strain, lowest proportion in the sample ($\pi_2 = 0.05$), is also the most stable over time ($\sigma^2_{\beta_r 2} = 0.19$). This strain has mostly segregating sites in all occasions and is represented by the almost completely white rows in Figure 5.8.

On the other hand, the other two strains are much bigger ($\pi_1 = 0.35$ and $\pi_3 = 0.60$)and have way higher levels of variation over time ($\sigma^2_{\beta_r 1} = 1.46$ and $\sigma^2_{\beta_r 3} = 1.6$). Although having somewhat parallel patterns, the estimates for $\alpha_r + \beta_r j$ for the biggest strain ($pi_3 = 0.6$) cross the cut points μ a number of times, for instance at occasions 3, 5 and 9. This reflects a higher variability on the ordinal responses over time on this most prevalent strain. The remaining strain ($\pi_1 = 0.35$) mostly stays at one end of the ordinal scale but is more likely to move around occasions 2 and 6. Importantly, the estimated variances of the random effects by cluster, $\sigma^2_{\beta_r} = (1.46, 0.19, 1.6)$ are very different and thus highlight the importance of including occassion-cluster interactions for this dataset.

Table 5.7: Posterior estimates for the model with $R = 3$ and $\beta_{rj} \sim$ Normal$(\beta_{rj-1}, \sigma^2_{\beta_r})$ for the infant gut data (84 parameters)

			95% Credible Interval		PSRF	
Par	Mean	SE	lower	upper	Point.est.	Upper.C.I.
μ_2	0.51	0.02	0.48	0.54	1.00	1.00
σ^2_μ	0.51	0.27	0.19	0.95	1.00	1.00
α_1	2.77	0.19	2.40	3.16	1.03	1.06
α_2	0.75	0.14	0.50	1.03	1.00	1.01
α_3	0.41	0.07	0.27	0.54	1.01	1.01
σ^2_α	1.43	0.50	0.70	2.37	1.00	1.00
β_{12}	-0.58	0.22	-1.00	-0.16	1.01	1.04
β_{13}	2.67	0.63	1.58	3.93	1.00	1.00
β_{14}	1.41	0.35	0.76	2.12	1.01	1.02
β_{15}	-0.04	0.25	-0.53	0.45	1.02	1.05

Continued on next page...

			95% Credible Interval		PSRF	
Par	Mean	SE	lower	upper	Point.est.	Upper.C.I.
β_{16}	-2.03	0.21	-2.44	-1.63	1.02	1.04
β_{17}	-1.74	0.20	-2.18	-1.39	1.02	1.05
β_{18}	-0.92	0.22	-1.35	-0.50	1.02	1.06
β_{19}	-0.58	0.23	-1.02	-0.11	1.02	1.04
β_{110}	1.33	0.36	0.64	2.08	1.00	1.01
β_{111}	3.83	0.66	2.57	5.10	1.00	1.00
β_{112}	4.25	0.79	2.75	5.79	1.00	1.01
β_{113}	4.36	0.86	2.78	6.04	1.00	1.00
β_{114}	3.50	0.65	2.23	4.76	1.00	1.01
β_{115}	4.82	0.88	3.22	6.58	1.00	1.01
β_{116}	4.40	0.93	2.69	6.21	1.00	1.00
β_{117}	3.74	0.66	2.50	5.06	1.00	1.01
β_{118}	2.99	0.51	1.99	3.99	1.00	1.01
β_{119}	2.73	0.49	1.77	3.71	1.01	1.02
β_{120}	4.72	0.93	3.00	6.57	1.00	1.01
β_{121}	5.32	1.04	3.49	7.49	1.00	1.00
β_{122}	4.84	0.92	3.09	6.64	1.00	1.00
β_{123}	4.80	0.96	3.07	6.74	1.00	1.00
β_{124}	4.51	0.83	3.02	6.20	1.00	1.01
β_{125}	3.98	0.84	2.43	5.65	1.01	1.01
β_{22}	-0.04	0.14	-0.30	0.26	1.00	1.00
β_{23}	-0.12	0.17	-0.45	0.22	1.00	1.01
β_{24}	-0.07	0.18	-0.41	0.27	1.00	1.01
β_{25}	0.04	0.18	-0.29	0.41	1.00	1.00
β_{26}	0.14	0.19	-0.22	0.51	1.00	1.00
β_{27}	0.07	0.19	-0.30	0.43	1.00	1.01
β_{28}	0.13	0.19	-0.23	0.50	1.00	1.01
β_{29}	0.05	0.19	-0.33	0.38	1.00	1.01
β_{210}	-0.11	0.18	-0.48	0.23	1.00	1.01
β_{211}	-0.29	0.18	-0.66	0.03	1.00	1.01
β_{212}	-0.38	0.18	-0.75	-0.05	1.00	1.01
β_{213}	-0.40	0.18	-0.76	-0.06	1.00	1.01
β_{214}	-0.40	0.18	-0.74	-0.05	1.00	1.01
β_{215}	-0.41	0.18	-0.74	-0.06	1.00	1.01
β_{216}	-0.43	0.18	-0.77	-0.08	1.00	1.01

Continued on next page...

			95% Credible Interval		PSRF	
Par	Mean	SE	lower	upper	Point.est.	Upper.C.I.
β_{217}	-0.44	0.18	-0.79	-0.11	1.00	1.01
β_{218}	-0.44	0.18	-0.78	-0.10	1.00	1.01
β_{219}	-0.43	0.18	-0.79	-0.09	1.00	1.01
β_{220}	-0.40	0.17	-0.74	-0.05	1.00	1.01
β_{221}	-0.40	0.18	-0.77	-0.08	1.00	1.00
β_{222}	-0.36	0.18	-0.70	-0.03	1.00	1.01
β_{223}	-0.37	0.18	-0.72	-0.04	1.00	1.01
β_{224}	-0.39	0.18	-0.75	-0.06	1.00	1.01
β_{225}	-0.40	0.19	-0.76	-0.02	1.00	1.00
β_{32}	-0.29	0.09	-0.48	-0.12	1.01	1.02
β_{33}	2.79	0.14	2.52	3.06	1.00	1.00
β_{34}	1.76	0.10	1.56	1.96	1.00	1.01
β_{35}	0.02	0.09	-0.17	0.19	1.01	1.02
β_{36}	-1.50	0.10	-1.69	-1.30	1.00	1.01
β_{37}	-1.55	0.10	-1.74	-1.34	1.00	1.01
β_{38}	-1.00	0.10	-1.18	-0.80	1.00	1.01
β_{39}	-0.50	0.09	-0.69	-0.32	1.00	1.01
β_{310}	0.69	0.09	0.52	0.88	1.00	1.01
β_{311}	5.29	0.37	4.61	6.07	1.00	1.00
β_{312}	7.10	0.83	5.56	8.74	1.00	1.00
β_{313}	7.33	0.99	5.67	9.36	1.00	1.00
β_{314}	7.70	0.96	5.97	9.61	1.00	1.00
β_{315}	8.15	1.04	6.29	10.21	1.00	1.00
β_{316}	6.78	0.80	5.40	8.40	1.00	1.00
β_{317}	7.31	0.83	5.77	8.94	1.00	1.00
β_{318}	7.43	0.94	5.81	9.34	1.00	1.00
β_{319}	6.53	0.65	5.26	7.78	1.00	1.00
β_{320}	5.73	0.46	4.86	6.65	1.00	1.00
β_{321}	8.13	1.10	6.19	10.25	1.00	1.00
β_{322}	9.01	1.34	6.58	11.50	1.00	1.00
β_{323}	9.00	1.47	6.57	12.04	1.00	1.00
β_{324}	8.70	1.26	6.37	11.11	1.00	1.00
β_{325}	7.34	1.15	5.54	9.66	1.00	1.00
$\sigma^2_{\beta 1}$	1.46	0.29	0.95	2.04	1.00	1.00
$\sigma^2_{\beta 2}$	0.19	0.04	0.12	0.27	1.00	1.01

Continued on next page...

			95% Credible Interval		PSRF	
Par	Mean	SE	lower	upper	Point.est.	Upper.C.I.
$\sigma^2_{\beta 3}$	1.69	0.30	1.21	2.34	1.00	1.01
π_1	0.35	0.01	0.33	0.38	1.00	1.01
π_2	0.05	0.00	0.04	0.05	1.00	1.00
π_3	0.60	0.01	0.57	0.63	1.00	1.01
log-like	-16763.02	5.33	-16773.96	-16753.23	1.00	1.00
log-post	-16860.69	7.14	-16875.51	-16847.71	1.00	1.00

In the next chapter, we will use latent transitional terms into the formulation of the POM. We called the resulting models, *data dependent* to emphasize its dependence to observed lagged responses, as opposed to the ones presented in this chapter where the linear predictor depended on latent lagged variables.

Appendix: Posterior estimates for the misspecified model

Table 5.8: Convergence results with model with prior $\beta_{rj} \sim \text{Normal}(0, \sigma^2_{\beta_r})$ when true model has $\beta_{rj} \sim N(\beta_{rj-1}, \sigma^2_r)$

				95% Credible Interval		PSRF	
Par	True	Mean	SE	lower	upper	Point.est.	Upper.C.I.
μ_2	0.85	0.86	0.04	0.79	0.94	1.00	1.00
μ_3	3.04	3.08	0.07	2.95	3.23	1.00	1.00
μ_4	3.89	3.90	0.09	3.73	4.06	1.00	1.00
σ^2_μ		1.54	0.44	0.83	2.34	1.00	1.00
α_1	-3.00	-2.79	0.18	-3.15	-2.46	1.00	1.00
α_2	1.00	0.78	0.15	0.47	1.06	1.00	1.01
α_3	3.00	2.90	0.15	2.61	3.20	1.00	1.00
σ^2_α		1.90	0.60	0.96	3.04	1.00	1.00
β_{12}	-0.08	0.08	0.28	-0.47	0.62	1.00	1.01
β_{13}	-0.05	-0.23	0.31	-0.84	0.36	1.01	1.03
β_{14}	0.05	-0.03	0.28	-0.60	0.51	1.00	1.00
β_{15}	0.06	-0.39	0.32	-1.03	0.22	1.00	1.02
β_{16}	0.19	-0.23	0.30	-0.80	0.40	1.01	1.01
β_{17}	0.27	-0.22	0.31	-0.84	0.38	1.00	1.01
β_{18}	0.44	0.09	0.27	-0.41	0.66	1.00	1.00
β_{19}	0.54	0.56	0.27	0.05	1.09	1.00	1.01
β_{110}	0.41	-0.14	0.30	-0.80	0.40	1.00	1.00
β_{22}	-0.71	-0.54	0.19	-0.90	-0.16	1.00	1.01
β_{23}	-0.72	-0.65	0.19	-1.04	-0.27	1.01	1.02
β_{24}	-1.36	-0.98	0.20	-1.36	-0.58	1.00	1.01
β_{25}	-1.42	-1.29	0.21	-1.67	-0.85	1.00	1.01
β_{26}	-1.80	-1.42	0.21	-1.83	-1.00	1.00	1.01
β_{27}	-1.28	-0.95	0.20	-1.33	-0.54	1.00	1.02
β_{28}	-1.15	-1.01	0.20	-1.42	-0.62	1.00	1.01
β_{29}	-0.88	-0.68	0.20	-1.06	-0.30	1.00	1.01
β_{210}	-0.59	-0.40	0.20	-0.78	-0.02	1.00	1.01
β_{32}	0.05	0.00	0.18	-0.36	0.32	1.00	1.00

Continued on next page...

				95% Credible Interval		PSRF	
Par	True	Mean	SE	lower	upper	Point.est.	Upper.C.I.
β_{33}	-0.31	-0.23	0.18	-0.59	0.12	1.00	1.00
β_{34}	-0.63	-0.34	0.19	-0.72	0.02	1.00	1.00
β_{35}	-0.68	-0.47	0.19	-0.84	-0.09	1.00	1.00
β_{36}	-0.48	-0.30	0.19	-0.68	0.06	1.00	1.00
β_{37}	-0.97	-0.58	0.19	-0.97	-0.20	1.00	1.00
β_{38}	-0.88	-0.70	0.20	-1.06	-0.30	1.00	1.00
β_{39}	-1.03	-0.96	0.20	-1.34	-0.54	1.00	1.00
β_{310}	-0.79	-0.57	0.19	-0.94	-0.19	1.00	1.00
$\sigma^2_{\beta 1}$	0.25	0.45	0.15	0.22	0.76	1.03	1.09
$\sigma^2_{\beta 2}$	0.50	0.95	0.25	0.54	1.45	1.01	1.04
$\sigma^2_{\beta 3}$	0.25	0.61	0.18	0.31	0.98	1.00	1.00
π_1	0.33	0.34	0.01	0.31	0.37	1.00	1.00
π_2	0.33	0.33	0.01	0.30	0.36	1.00	1.00
π_3	0.33	0.33	0.01	0.30	0.36	1.00	1.00
log-like	-6158.29	-6164.14	4.79	-6173.32	-6154.75	1.00	1.00

Chapter 6

Data dependent models

6.1 Latent transitional models

In this chapter, we develop a latent transitional model that is an extension of the Proportional Odds model (McCullagh 1980) and includes a mixture of first-order transitional terms to account for the repeated measures correlation. The proposed model-based clustering model thus includes both observed (lagged response) and latent covariates (cluster membership). As noted in the literature review, these are also known as Markov transition or latent transition models and have been proposed by several authors (Frydman 2005, Pamminger et al. 2010, Frühwirth-Schnatter et al. 2012, Cheon et al. 2014). Unlike these however, the approach developed in this chapter models explicitly the ordinal nature of the response by using cumulative distribution functions while also allowing for time-varying transition probabilities. We estimate the model within a Bayesian setting using a MCMC scheme and block-wise Metropolis-Hastings sampling.

6.2 Model

Let Y be an ordinal response with q levels measured over n subjects on p occasions, with indexes i, j, k for subjects, occasions, and ordinal levels, respectively. Suppose that subjects come from latent cluster r with probability π_r ($r = 1, \ldots, R$; $\sum_{r=1}^{R} \pi_r = 1$) and let $\theta_{rk'kj} = P(y_{ij} = k | i \in r, y_{i(j-1)} = k')$. We extend the POM (McCullagh 1980) to include a latent cluster, a previous response and occasion effects. We model the cumulative probability of each ordinal outcome as

$$\text{Logit}[P(y_{ij} \leq k | i \in r, y_{i(j-1)} = k')] = \mu_k - \alpha_r - \beta_{k'} - \gamma_j \tag{6.1}$$

or alternatively:

$$\begin{aligned} y_{ij} \mid i \in r, y_{i(j-1)=k'} &\sim \text{Categorical}_q(\theta_{rk'.j}) \ , \ \sum_{k=1}^{q} \theta_{rk'kj} = 1 \\ \theta_{rk'kj} &= \frac{1}{1 + e^{-(\mu_k - \alpha_r - \beta_{k'} - \gamma_j)}} - \frac{1}{1 + e^{-(\mu_{k-1} - \alpha_r - \beta_{k'} - \gamma_j)}} \\ & \\ & i = 1, \ldots, n \\ & j = 2, \ldots, p \\ & r = 1, \ldots, R \\ & \mu_{k-1} < \mu_k; \ k = 0, \ldots, q; \ \mu_0 = -\infty; \mu_1 = 0 \text{ and } \mu_q = \infty \\ & \beta_q = 0; k' = 1, \ldots, q \\ & \gamma_2 = 0; j = 2, \ldots, p \end{aligned} \tag{6.2}$$

That is, each response y_{ij} is the realization of a categorical distribution with probabilities $\theta_{rk'1j}, \ldots \theta_{rk'qj}$. Notice that the linear predictor for the probability $\theta_{rk'kj}$ contains both observed (previous response $y_{i(j-1)}$ and occasion j) and unobserved covariates (cluster membership for subject i). The parameter μ_k is the cut point for each ordinal category, with the same

parametrisation used in Chapter 5 with $\mu_1 = 0$. The parameter α_r is the effect of the latent cluster r, the parameter $\beta_{k'}$ the effect of having an outcome k' at the previous occasion and the parameter γ_j the effect of occasion j. The choice of a negative sign preceding α_r, γ and $\beta_{k'}$ implies that increases in these coefficients increase the probability of observing outcomes in the upper-end of the ordinal scale (closer to q than to 1). Note finally that we do not model the first response and instead condition on its value.

6.2.1 Likelihood

Given the dependence on the previous outcome, we can factorize the likelihood to separate the contribution of the first occasion, $Y = (Y_1, \widetilde{Y})$ where $Y_1 = \{Y_{i1}, \forall i\}$. Assuming independence over the rows and independence over the columns conditional on the rows and the previous response, the model's likelihood for the transitions ($j \geq 2$) becomes

$$L(\widetilde{Y}|\mu, \alpha, \beta, \pi, y_{i(j-1)}) = \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=2}^{p} \prod_{k'=1}^{q} \prod_{k=1}^{q} \theta_{rk'kj}^{I(y_{ij}=k, y_{i(j-1)}=k')} \tag{6.3}$$

where: $\pi_r \geq 0$, $\sum_{r=1}^{R} \pi_r = 1$ and $I(.)$ is an indicator function equal to 1 if the argument is true, and 0 otherwise.

6.2.2 Bayesian Estimation

The model is completed with the following weakly informative priors:

$$
\begin{aligned}
\mu \mid \sigma_\mu^2 &\overset{iid}{\sim} \text{OS[Normal}(0, \sigma_\mu^2)],\ \mu_k > \mu_{k-1};\ k = 0 \ldots q;\ \mu_0 = -\infty; \mu_1 = 0; \mu_q = \infty \\
\alpha_r \mid \sigma_\alpha^2 &\overset{iid}{\sim} \text{Normal}(0, \sigma_\alpha^2),\ r = 1, \ldots, R; \\
\beta_{k'} \mid \sigma_\beta^2 &\overset{iid}{\sim} \text{Normal}(0, \sigma_\beta^2),\ k' = 1, \ldots, q;\ \beta_q = 0 \\
\gamma_j \mid \sigma_\gamma^2 &\overset{iid}{\sim} \text{Normal}(0, \sigma_\gamma^2),\ j = 2, \ldots, p;\ \gamma_2 = 0 \\
\sigma_\mu^2 &\sim \text{Inverse Gamma}(a_\mu, b_\mu) \\
\sigma_\alpha^2 &\sim \text{Inverse Gamma}(a_\alpha, b_\alpha) \\
\sigma_\beta^2 &\sim \text{Inverse Gamma}(a_\beta, b_\beta) \\
\sigma_\gamma^2 &\sim \text{Inverse Gamma}(a_\gamma, b_\gamma) \\
\pi &\sim \text{Dirichlet}(\psi),\ r = 1 \ldots R
\end{aligned}
\tag{6.4}
$$

where OS=Order Statistics and the hyperparameters are set to: $a_\mu = a_\alpha = a_\beta = a_\gamma = 4$, $b_\mu = b_\alpha = b_\beta, b_\gamma = 0.5$, and $\psi = 1.5$.

In words, we assign Truncated Normal priors for the cut-off points μ, Normal priors centered on zero and with an unknown variance for α, γ and β, a Dirichlet prior for the mixing probabilities π, and Inverse Gamma priors for the unknown variances σ_μ^2, σ_α^2, σ_γ^2 and σ_β^2. Figure 6.1 shows a graphical representation of the model and all priors.

Given the likehood (equation 6.3), the posterior distributions for the model parameters are not available in closed form. To perform the posterior computation, we use a Markov chain Monte Carlo (MCMC) sampling scheme. In particular, we use a Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970) with random walk proposals to sample blocks of parameters separately. With the exception of the transitional term β_k' and column effects β_j, the model in (6.1) has a similar parameter vector than the latent random effects model in Chapter 5. We therefore use a similar Metropolis-Hasting scheme to simulate the target distribution. Proposals and MH ratios could be found in the Appendix at the end of the chapter.

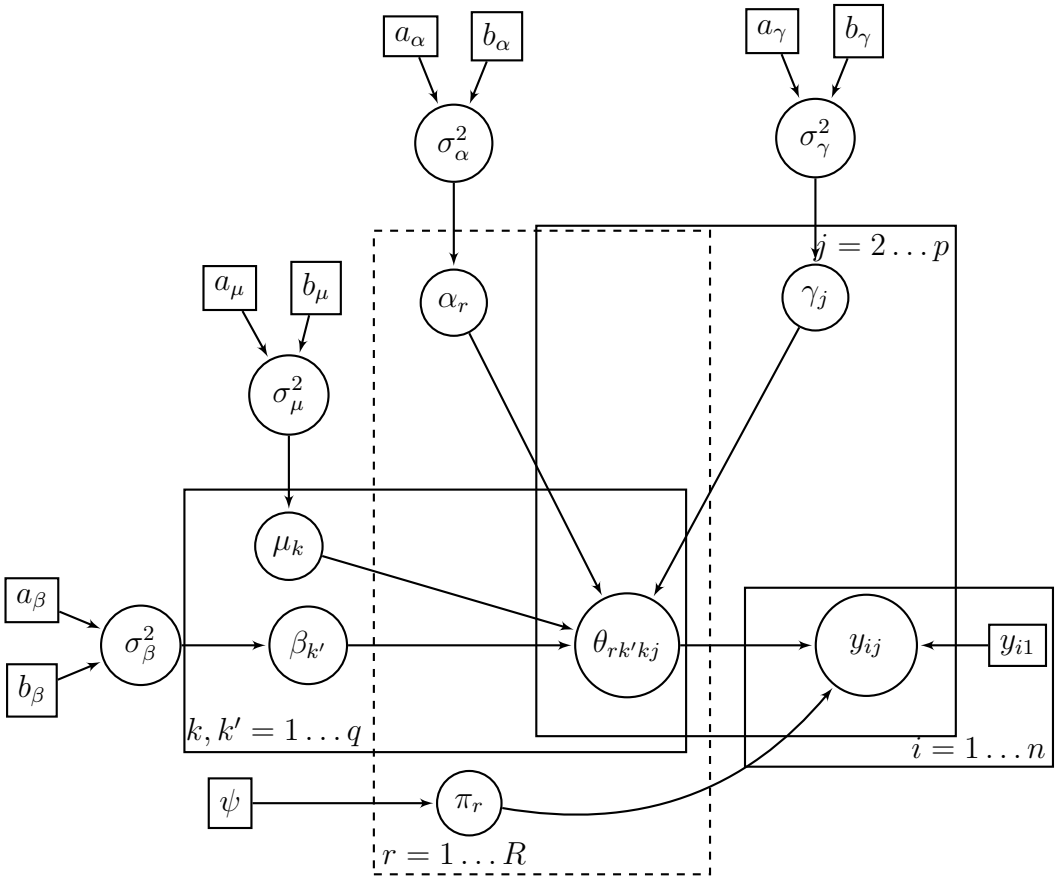


Figure 6.1: Graphical representation of the model (6.2) and priors (6.4).

6.2.3 MCMC Convergence

Similarly to the previous chapter, we use the potential scale reduction factor (PSRF), or Gelman-Rubin convergence diagnostic (Gelman & Rubin 1992), to assess convergence of the MCMC chain. Lack of convergence is indicated by PSRF values much higher than one. As remarked before, the Gelman-Rubin convergence diagnostic uses parallel chains to monitor convergence, between-chain and within-chain variances, and thus is very helpful in detecting where chains have, or have not, converged to the same mode.

6.2.4 Model Comparison

We also use the Widely Applicable Information Criterion (WAIC) (Watanabe 2009) for model comparison in this chapter. See section 5.4 in Chapter 5 for formulas to calculate the WAIC and its components: log predictive density (LPD) and number of effective parameters (p).

Additionally, we also look at the entropy of the resulting clustering. As such this is a different model comparison criterion, since entropy measures focus on the degree of separation of the mixture components and not the predictive density like the WAIC, and AIC in the Frequentist approaches. The entropy of the classification probabilities thus provides an alternative way to compare among candidate models.

Entropy for classification probabilities

In all mixture models, each observation has some probability of coming from a mixture component. Denoting s as an MCMC draw and keeping the notation used earlier in the chapter, we compute classification probabilities $\hat{z}_{ir}$ as

.

$$\hat{z}_{ir} = E_s[z_{ir}^s | Y, \phi^s, \pi^s] = \frac{\pi_r^s \prod_{j=2}^{p} \prod_{k'=1}^{q} \prod_{k=1}^{q} \theta_{rk'kj}^{s,I(y_{ij}=k)}}{\sum_{a=1}^{R} \pi_a^s \prod_{j=2}^{p} \prod_{k'=1}^{q} \prod_{k=1}^{q} \theta_{ak'kj}^{s,I(y_{ij}=k)}} \tag{6.5}$$

That is, $\hat{z}_{ir}$ is the posterior mean of the classification probabilities z_{ir}^s over the MCMC chain. The latter is in turn calculated using the parameters ϕ^s, π^s at each s. Note that, for each s this is the same formulae in the E-step of the EM algorithm used before (equation 3.20 in chapter 3).

Given these classification probabilities $\hat{z}_{ir}$ we calculate its associated entropy (EN) according to

$$\text{EN} = \sum_{i=1}^{n} \sum_{r=1}^{R} \hat{z}_{ir} \log(\hat{z}_{ir}) \tag{6.6}$$

Importantly, the entropy for a given model is also its estimated Kullback-Leibler (KL) distance from the true model, i.e. a measure of the distance between these two probability distributions. In the context of mixture models, it also can be interpreted as a proxy for the degree of separation of the mixture components. That is, mixtures with well-separated components will have lower entropy, high estimated classification probabilities and as a result crisp allocation of subjects to clusters. On the contrary, the reverse is also true, not so well-separated mixture components will be associated with higher entropy and fuzzy clustering.

Note also that for mixture models, the maximum value of the entropy increases with the number of mixture components (R). That is, the entropy of the least informative classification probabilities changes according to R and it is equal to $\hat{z}_{ir} = 1/R, \forall r, i$. With two components, for instance, the least informative $\hat{z}_{ir} = (0.5, 0.5), \forall i$. Such classification probabilities are the ones we expect if would randomly allocate subjects i to clusters r.

In light of the above, we also calculate the ratio of the estimated entropy of a mixture model and the entropy of a model, with the same n and R, that has the least informative classification probabilities, i.e. the entropy of a model that allocates observations to clusters randomly. We call this quantity *relative entropy* and informally interpret it as a proxy for the KL distance of the model from chance. That is, the relative entropy provides a proxy of how far the estimated clustering is from a random classification.

6.3 Simulations

We first validate the model using simulated data. We simulate a medium size dataset with 600 rows, 15 columns and five ordinal categories from a mixture with three components with equal proportions. This model has 29 parameters. We used five chains with over-dispersed starting values and ran 3.6 million iterations for each chain. Discarding the initial 25% draws

as burn-in and thinning these chains by 4000, we used for inference 3375 MCMC draws.

Table 6.1 shows the true values of model parameters as well as summaries (mean, SE and credible intervals) of the estimated marginal posteriors. This table also shows the Gelman-Rubin convergence diagnostic (point estimate and upper confidence interval for the PSRF) which are all close to 1 and well below the threshold value of no convergence of 1.2. As it can be seen, the model is able to able to recover all parameters as the 95% credible intervals for all 29 parameters include the true values. The table also shows that this also holds true for the log-likelihood and the joint posterior.

In addition to that, Figure 6.2 shows the posterior classification probabilities by the estimated model. All classification probabilities are very close to one (0.997 overall median) which implies that the clustering provided by the model is crisp, ie individuals are assigned to an estimated cluster with very high probability.

With regard to model comparison, Table 6.2 present the WAIC and entropy measures (median and relative) for models with $R = 1 \dots 4$. All these measures, choose the true model with $R = 3$ which has both the lowest entropy (-59) and relative entropy (0.09). This reassures us that the proposed measures can identify the most appropriate model among several candidates model with different number of mixture components.

Table 6.1: Summary statistics for the marginal posteriors and Gelman-Rubin convergence diagnostic (PSRF) for the simulated data

				95% Credible Interval		PSRF	
Par	True	Mean	SE	lower	upper	Point.est.	Upper.C.I.
μ_2	1.96	1.96	0.05	1.86	2.06	1.00	1.00
μ_3	3.58	3.56	0.06	3.45	3.69	1.00	1.00
μ_4	5.55	5.55	0.08	5.40	5.71	1.00	1.00
σ^2_μ	-	3.07	0.99	1.62	4.99	1.01	1.02
α_1	2.77	2.87	0.13	2.63	3.11	1.00	1.00
α_2	0.96	1.02	0.13	0.75	1.27	1.00	1.02
α_3	4.77	4.86	0.12	4.66	5.08	1.00	1.00
σ^2_α	2.00	2.48	0.77	1.32	3.91	1.00	1.00
β_1	-1.67	-1.72	0.11	-1.92	-1.51	1.00	1.01
β_2	-1.44	-1.48	0.09	-1.67	-1.31	1.00	1.00
β_3	-0.96	-1.07	0.09	-1.22	-0.89	1.00	1.00
β_4	-0.96	-1.04	0.08	-1.22	-0.90	1.00	1.00
σ^2_β	0.50	1.22	0.38	0.61	1.87	1.00	1.00
γ_3	1.01	0.84	0.11	0.63	1.05	1.00	1.02
γ_4	1.25	1.09	0.11	0.88	1.31	1.00	1.01
γ_5	1.22	1.10	0.11	0.89	1.31	1.00	1.01
γ_6	1.96	1.88	0.11	1.66	2.09	1.00	1.00
γ_7	3.21	3.11	0.12	2.89	3.35	1.01	1.02
γ_8	1.32	1.14	0.11	0.91	1.35	1.00	1.01
γ_9	1.76	1.62	0.11	1.42	1.84	1.01	1.04
γ_{10}	-0.26	-0.34	0.11	-0.56	-0.15	1.00	1.02
γ_{11}	1.34	1.28	0.11	1.06	1.48	1.00	1.01
γ_{12}	1.73	1.65	0.11	1.44	1.85	1.00	1.01
γ_{13}	0.88	0.81	0.11	0.59	1.03	1.00	1.00
γ_{14}	1.41	1.29	0.11	1.06	1.48	1.00	1.00
γ_{15}	0.75	0.85	0.11	0.64	1.09	1.00	1.00
σ^2_γ	1.00	1.37	0.27	0.96	1.96	1.00	1.01
π_1	0.33	0.35	0.02	0.32	0.38	1.00	1.00
π_2	0.33	0.35	0.02	0.32	0.38	1.00	1.00
π_3	0.33	0.30	0.01	0.27	0.33	1.00	1.01
log-like	-10649.99	-10651.34	3.80	-10658.36	-10644.02	1.00	1.02
log-post	-10743.65	-10712.52	3.95	-10720.66	-10705.70	1.00	1.02

R	pars	LPD	p_{WAIC1}	WAIC_1	p_{WAIC2}	WAIC_2	Median Entropy (A)	Max Entropy for R mixture (B)	Relative Entropy (A/B)
1	24	22669.1	22.5	22714.0	22.5	22714.1	-	-	-
2	27	21740.3	30.0	21800.3	30.1	21800.5	-60.8	-415.9	0.15
3	29	21276.7	25.9	21328.6	26.0	21328.8	-59.0	-659.2	0.09
4	31	21275.0	27.1	21329.1	27.1	21329.1	-161.8	-831.8	0.19

Table 6.2: Bayesian model comparison using WAIC for the simulated data

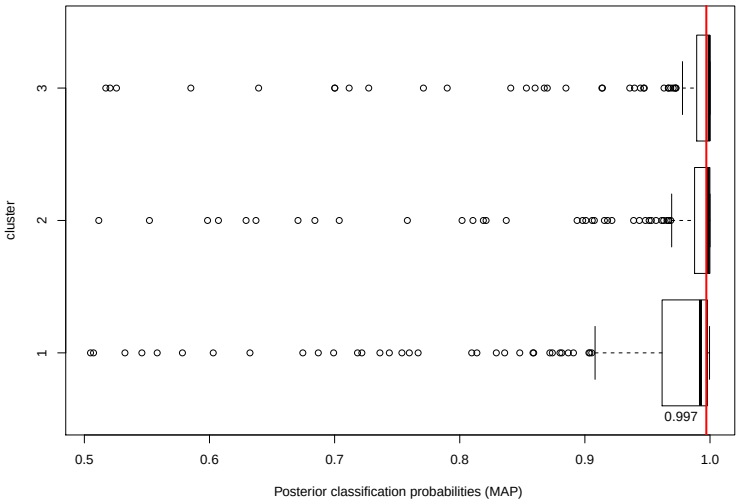


Figure 6.2: Posterior classification probabilities and entropy distribution for the simulated data

6.4 Case Study: 2001-2011 self reported health status from HILDA

In this section, we fit the model to self reported health status (SRHS) from the HILDA dataset. In short, SRHS is an ordinal variable with five levels ("Poor", "Fair", "Good", "Very Good" and "Excellent") measured for 4660 individuals over 2001-2011. Detailed information about this dataset can be found in chapter 2. We used five chains with over dispersed starting values and ran 3.6 million iterations for each chain. Discarding the initial 25% draws as burn-in and thinning these chains by 8000, we used for inference 1690 MCMC draws.

Table 6.3: Bayesian model comparison using WAIC for the HILDA dataset

R	pars	LPD	p_{WAIC1}	WAIC_1	p_{WAIC2}	WAIC_2	Median Entropy (A)	Max Entropy for R mixture (B)	Relative Entropy (A/B)
1	20	100683.6	26.2	100736.1	26.4	100736.4	-	-	-
2	23	87249.4	28.2	87305.9	28.3	87305.9	-551.3	-3230.1	0.171
3	25	85998.3	29.6	86057.4	29.6	86057.5	-811.6	-5119.5	0.159
4	27	85021.6	29.9	85081.4	30.0	85081.5	-999.2	-6460.1	0.155
5	29	84771.1	32.3	84835.6	32.3	84835.8	-1131.6	-7500.0	0.151
6	31	84599.7	32.2	84664.1	32.3	84664.3	-1867.0	-8349.6	0.224

Table 6.3 shows the results of model comparison. For each fitted model it presents: number of clusters (R), total number of parameters (Pars),the two versions of the WAIC, WAIC2 and their corresponding components (LPD, p, p_2). Both versions of the WAIC suggest the same conclusion, the model with six components seems to provide the best fit. In contrast to that, the relative entropy is the lowest in a model with $R = 5$. Guided by parsimony we choose the latter model with 5 mixture components for the HILDA dataset.

As in the previous chapter, we post-process the MCMC chains using the Stephens relabelling algorithm (Stephens 2000) to remove label switch-

Table 6.4: Summary statistics for the posteriors and Gelman-Rubin convergence diagnostic (PSRF) for R=5

				95% Credible Interval		PSRF	
Par	True	Mean	SE	lower	upper	Point.est.	Upper.C.I.
μ_2		3.73	0.06	3.62	3.86	1.00	1.00
μ_3		7.40	0.07	7.26	7.54	1.00	1.00
μ_4		11.28	0.08	11.13	11.44	1.00	1.00
σ^2_μ		6.13	2.15	3.02	9.89	1.00	1.00
α_1		3.80	0.13	3.55	4.06	1.00	1.01
α_2		5.97	0.09	5.78	6.15	1.00	1.00
α_3		8.12	0.09	7.96	8.30	1.00	1.01
α_4		10.20	0.09	10.02	10.37	1.00	1.00
α_5		12.47	0.09	12.29	12.65	1.00	1.00
σ^2_α		7.07	1.84	4.13	10.44	1.00	1.00
β_1		-4.75	0.11	-4.96	-4.53	1.00	1.00
β_2		-3.25	0.06	-3.37	-3.12	1.00	1.00
β_3		-2.07	0.05	-2.18	-1.97	1.00	1.00
β_4		-1.04	0.05	-1.13	-0.95	1.00	1.00
σ^2_β		2.61	0.84	1.37	4.15	1.00	1.00
γ_3		-0.04	0.04	-0.12	0.05	1.00	1.00
γ_4		-0.09	0.04	-0.17	-0.00	1.00	1.01
γ_5		-0.20	0.04	-0.29	-0.11	1.00	1.01
γ_6		-0.11	0.04	-0.19	-0.03	1.00	1.01
γ_7		-0.16	0.04	-0.25	-0.07	1.00	1.00
γ_8		-0.33	0.04	-0.42	-0.25	1.00	1.00
γ_9		0.02	0.04	-0.07	0.10	1.00	1.00
γ_{10}		-0.51	0.05	-0.61	-0.43	1.00	1.00
γ_{11}		-0.46	0.04	-0.55	-0.38	1.00	1.01
σ^2_γ		0.34	0.08	0.20	0.51	1.00	1.00
π_1		0.03	0.00	0.03	0.04	1.00	1.01
π_2		0.15	0.01	0.13	0.16	1.01	1.02
π_3		0.38	0.01	0.37	0.39	1.00	1.01
π_4		0.35	0.01	0.33	0.36	1.00	1.01
π_5		0.09	0.00	0.08	0.10	1.00	1.01
log-like		-42401.67	3.42	-42408.80	-42395.74	1.00	1.01
log-post		-42460.76	3.68	-42468.00	-42453.80	1.00	1.01

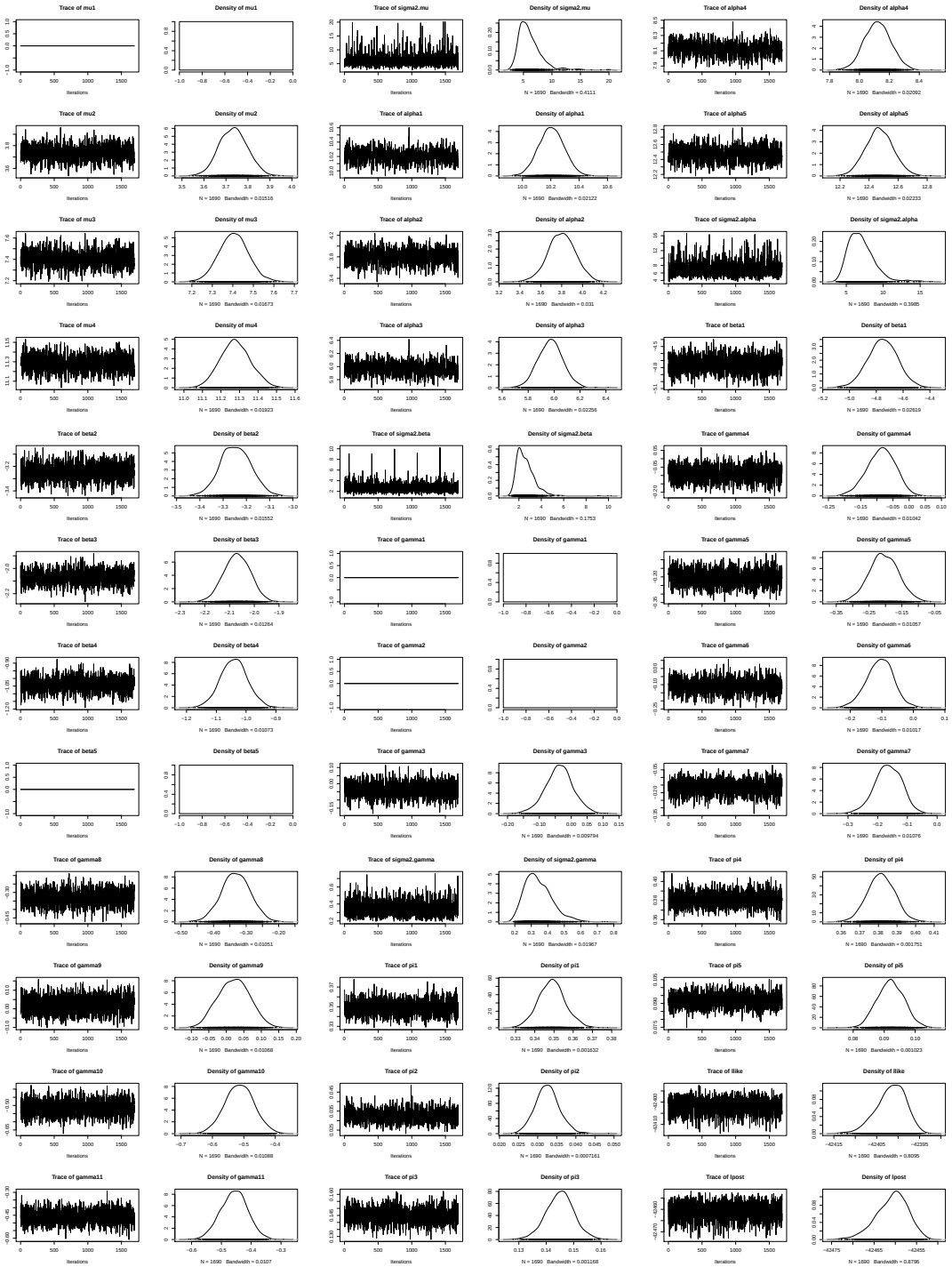
ing. Summary statistics for the posterior distributions, along with the PSRF convergence diagnostic, of the parameters of this model are shown in Table 6.4. Also, marginal posterior for all parameters and their corresponding traceplots are shown in Figure 6.3.

Next, we check for the fuzziness of the estimated cluster memberships probabilities for this model with five clusters. Figure 6.4 displays the co-clustering probabilities for the model in the original data (top panel) and ordered by cluster (bottom panel). Co-clustering probabilities are very high, around 80% in all clusters. In addition to that, Figure 6.5 shows the distribution of the estimated classification probabilities. As it can be seen, the membership probabilities are pretty high for all clusters with a overall median of 0.967%. In sum, a model with five clusters does not only provide the best fit among $R = 1 \dots 6$ but also provides a very crisp allocation of individuals to the estimated clusters.

What do these estimated clusters look like? Figure 6.6 shows heatmaps of the estimated transition probabilities by cluster $\hat{\theta}_{rk'kj}$ along with the cluster effects $\hat{\alpha}_r$ and proportions $\hat{\pi}_r$ for the selected model with five clusters. We calculate $\hat{\theta}_{rk'kj}$ evaluating (6.2) at the mean marginal posterior of each model parameter, i.e. $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$,γ and $\hat{\pi}$. This figure also includes the empirical transition probabilities for all data (top left corner), that is the year-to-year empirical probabilities averaged over 2001-2011. Starting with this latter heatmap, we could see that most of the transitions happened at or nearby the diagonal. Thus, individuals tend to report a very similar health status to the one they reported previously.

In contrast to that, the rest of heatmaps in Figure 6.6 show markedly different patterns. Firstly, the almost vertical shapes in these transition probabilities reflect the nearby or local nature of the transitions in this dataset. This means that regardless of their starting SRHS, individuals tend to move to nearby categories only. We already commented on the stability of the SRHS when looking at the raw data in chapter 2 but it interesting to see that the model is able to capture this characteristic.

Figure 6.3: Relabelled MCMC output for HILDA ($R = 5$)



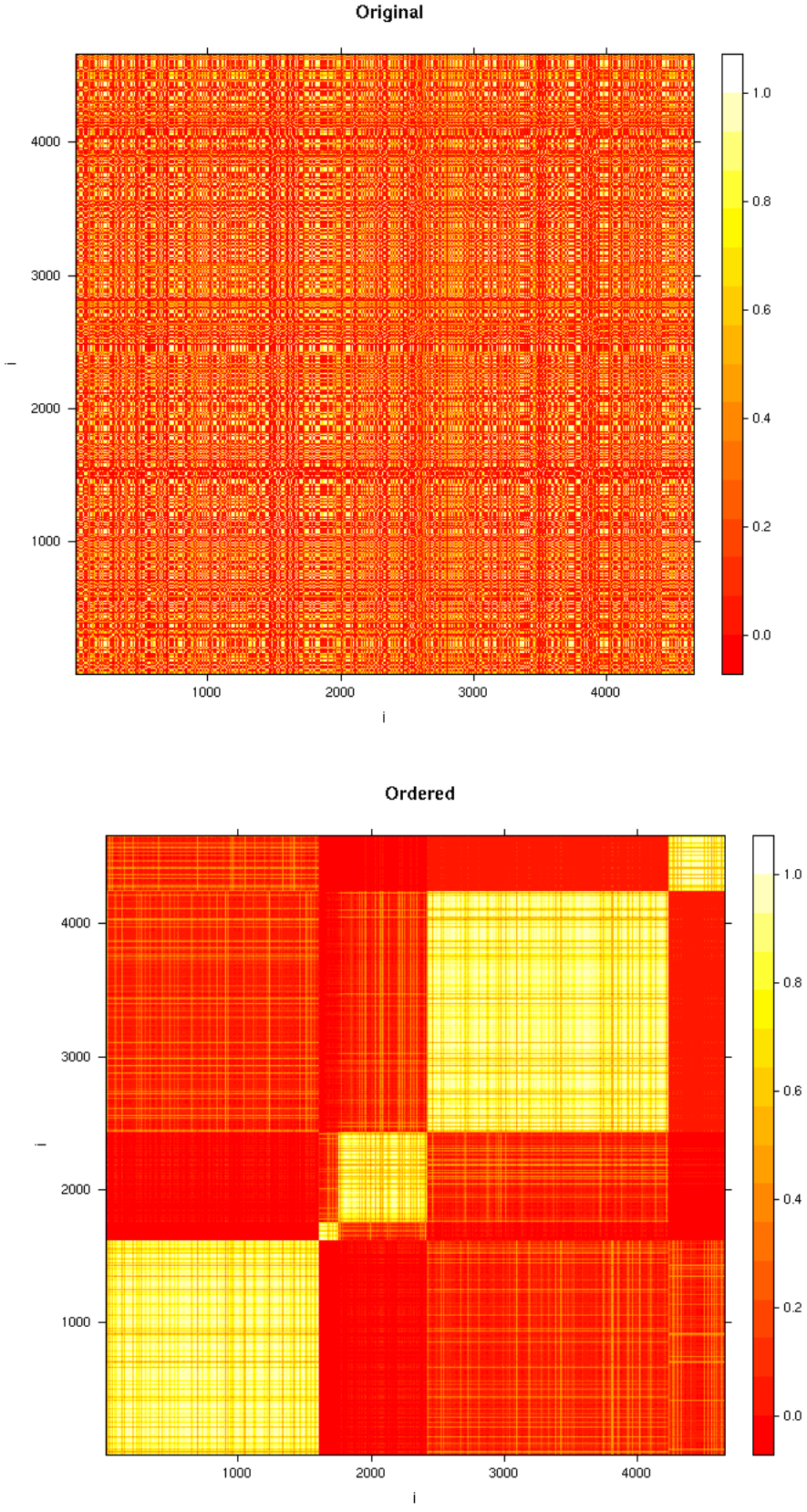


Figure 6.4: Co-clustering probabilities for HILDA $R = 5$

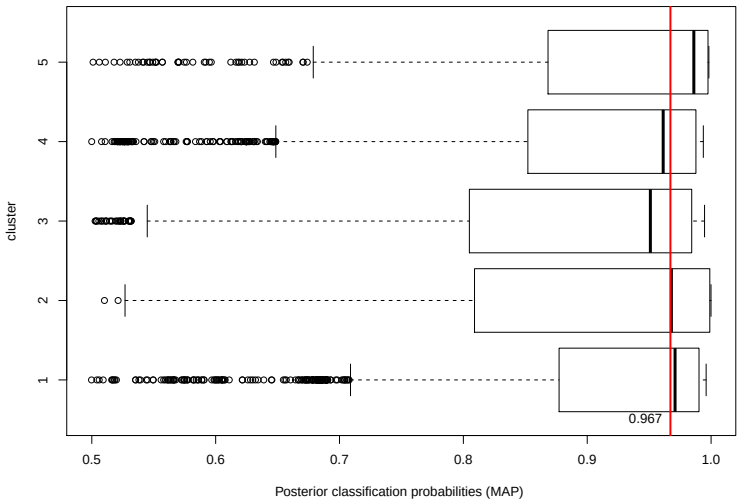


Figure 6.5: Distribution of the classification probabilities by cluster ($R = 5$) for HILDA. Overall median classification probability in red

Secondly, the estimated clusters are formed by respondents with similar SRHS levels. For example, individuals in cluster 1, $\hat{\alpha_1} = 3.8$ and 3% of the total, have transitions in the lower end of the ordinal category ("Poor" and "Fair"). Respondents in cluster 2, $\hat{\alpha_2} = 6$ and 5% of the total, are more neutral about their health status as they have transitions over the "Fair" and "Good" categories. Clusters 3 and 4, $\hat{\alpha_3} = 8$ and $\hat{\alpha_4} = 10.2$, are formed by individuals that are more positive about their SRHS with transitions over the middle-categories ("Fair", "Good" and "Very Good"). Importantly, these two latter clusters together are about half of the total ($\hat{\pi_3} = 0.38$ and $\hat{\pi_4} = 0.35$). Individuals in cluster 5, $\hat{\alpha_5} = 12.5$, represent about 9% of the total and are very positive about their SRHS, having only transitions between the "Good" and "Very Good" levels.

In addition to the above, the heatmaps also provide information on SRHS trends over time within each cluster. Transitions above the diagonal imply a worsening SRHS, higher probability of reporting a lower health

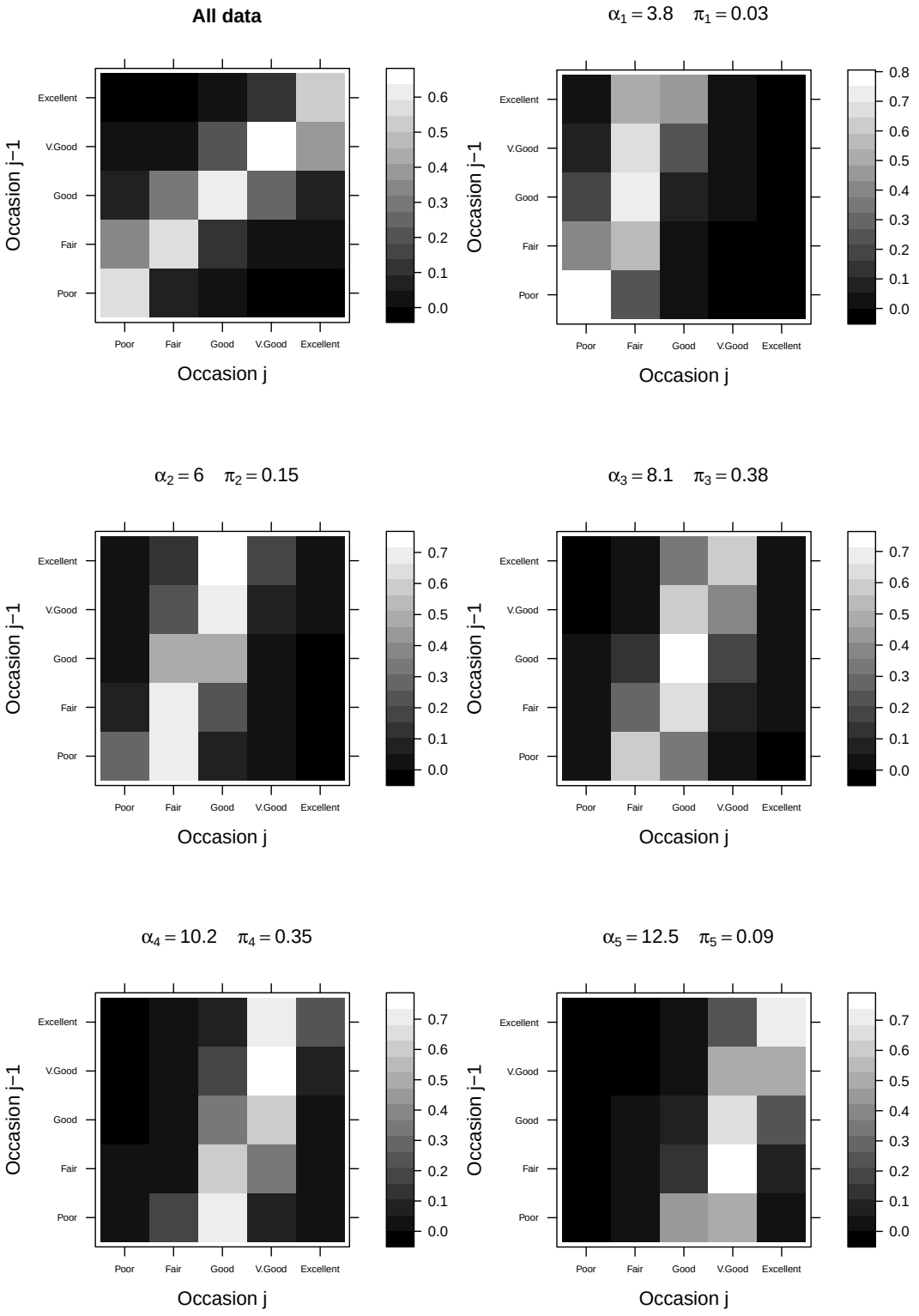


Figure 6.6: Estimated transition matrices by cluster over 2001-2011. Rows add to 1

status in the current period. Complementary, transitions below the diagonal imply an improving health status between occasions. In this regard, SRHS worsens in clusters 1 and 2 and improves in clusters 4 and 5. Cluster 3 has both types of respondents, i.e. estimated transitions in these clusters are both below and above the diagonal.

Finally, we present the distribution of SRHS by cluster and year in Figure 6.7. We also present the distribution of SRHS for all data for comparison. In this plot, we can see the different profiles of the SRHS distribution for each cluster. Clusters 1 and 2 are formed by people that have a negative/neutral perception of their health status (”Poor” and ”Fair” categories), people in cluster 3 position their SRHS exactly at the middle of the ordinal scale (”Good”), while respondents in the remaining clusters are way more positive about their health status. Cluster 5 represents, for instance, the extreme of positiveness of SRHS as respondents in these are extremely satisfied with their health status and have responses only in ”Very Good” and ”Excellent” categories. On the other hand, this figure also shows the temporal variations on the reported health status. Likewise in the heatmaps, SRHS gets worse over time in clusters 1 and 2 since these respondents tend to move towards the lower-end of the ordinal scale (”Poor”). On the other hand, SRHS improves over time in cluster 5 where the SRHS distribution moves towards the upper-end categories (”Very Good” and ”Excellent”). Lastly, in clusters 3 and 4 SRHS moves towards the middle category (”Good”) neither worsening nor improving but getting more ”central”.

Similarly to Chapter 5, the model proposed in the present chapter assumes that the number of mixture components is fixed, that is models are conditional on the number of mixture components R. We proceed to relax this assumption in the next chapter using a Dirichlet Process Mixture within a Bayesian Non-Parametric (BNP) framework. We show that this BNP model is tractable, i.e. is easily computed using MCMC standard methods.

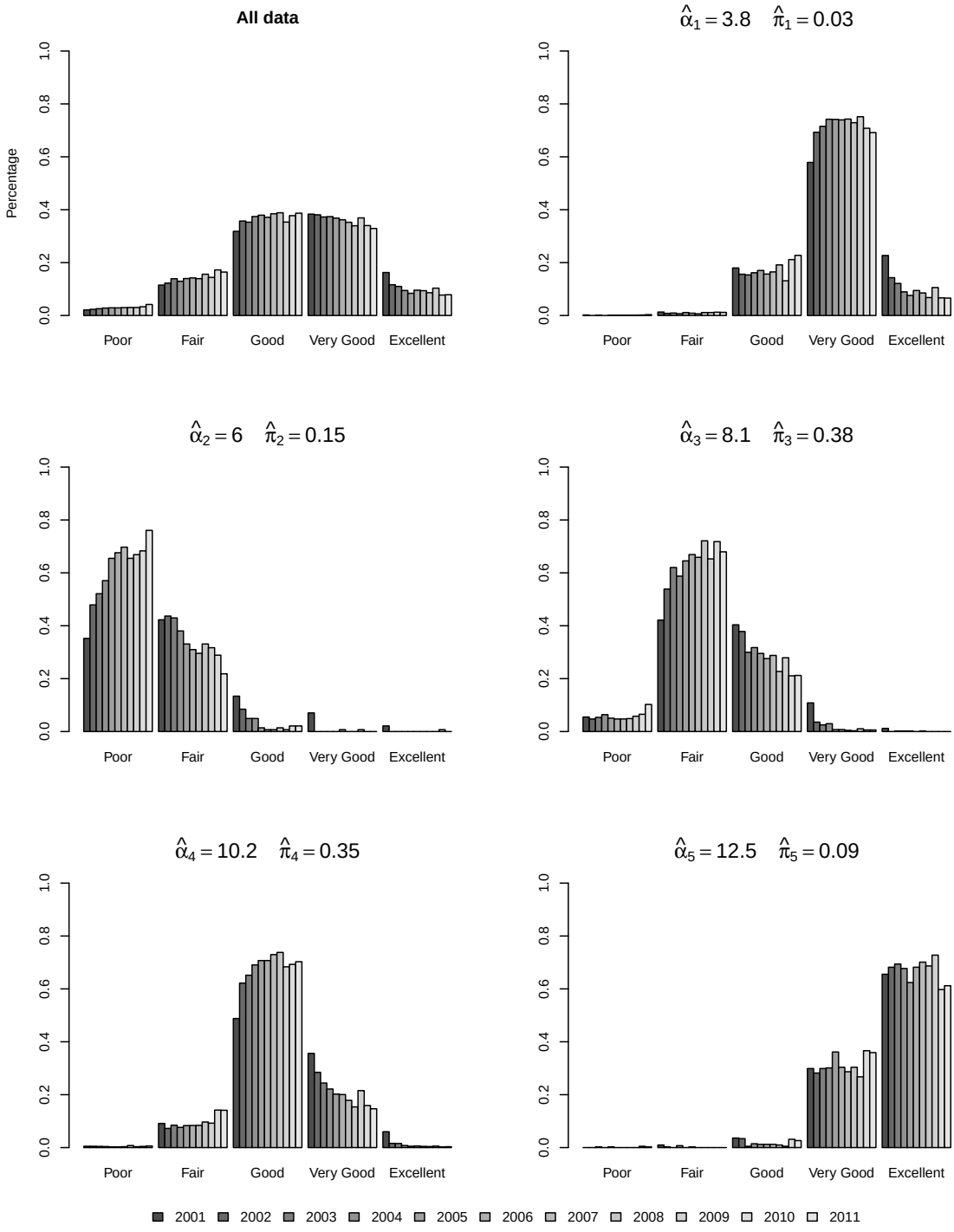


Figure 6.7: SRHS distribution by cluster over 2001-20011 for HILDA

Appendix: MH scheme for the latent transitional model

Proposals

Recall the parameter vector $\Omega = (\phi, \pi)$ where $\phi = (\mu, \alpha, \beta, \gamma, \sigma^2_\mu, \sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma)$ for the model developed in this chapter (equation 6.1). We first choose initial values for these parameters and then use random walk proposals to update them.

$$
\begin{aligned}
\mu'_k \mid \mu_{k-1}, \mu_k, \mu_{k+1} &\sim U[\,\max(\mu_k - \tau, \mu_{k-1}),\ \min(\mu_k + \tau, \mu_{k+1})] \\
&\qquad k = 2, \ldots, q-1,\ \mu_0 = -\infty, \mu_1 = 0,\ \mu_q = \infty \\
\alpha'_r \mid \alpha_r &\sim \text{Normal}(\alpha_r, \sigma^2_{\alpha p}) \qquad r = 1 \ldots R,\ \alpha_1 = 0 \\
\beta'_{k'} \mid \beta_{k'} &\sim \text{Normal}(\beta_{k'}, \sigma^2_{\beta p}) \qquad k' = 1 \ldots q,\ \beta_q = 0 \\
\gamma'_j \mid \gamma_j &\sim \text{Normal}(\gamma_j, \sigma^2_{\gamma p}) \qquad j = 3 \ldots p,\ \gamma_2 = 0 \\
\text{logit}(w') \mid \text{logit}(w) &\sim \text{Normal}(\text{logit}(w), \sigma^2_{\pi p}) \\
&\qquad w = \pi_{r1}/(\pi_{r1} + \pi_{r2})\ r1, r2 \in 1 \ldots R \\
&\qquad \pi'_{r1} = w'(\pi_{r1} + \pi_{r2}) \\
&\qquad \pi'_{r2} = (1 - w')(\pi_{r1} + \pi_{r2}) \\
\log(\sigma'^2_\mu) \mid \log(\sigma^2_\mu) &\sim \text{Normal}(\log(\sigma^2_\mu), \sigma^2_{\sigma\mu p}) \\
\log(\sigma'^2_\alpha) \mid \log(\sigma^2_\alpha) &\sim \text{Normal}(\log(\sigma^2_\alpha), \sigma^2_{\sigma\alpha p}) \\
\log(\sigma'^2_\beta) \mid \log(\sigma^2_\beta) &\sim \text{Normal}(\log(\sigma^2_\beta), \sigma^2_{\sigma\beta p}) \\
\log(\sigma'^2_\gamma) \mid \log(\sigma^2_\gamma) &\sim \text{Normal}(\log(\sigma^2_\gamma), \sigma^2_{\sigma\gamma p})
\end{aligned}
$$

We use the following step sizes: $\tau = 0.25$, $\sigma^2_{\alpha p} = 0.25$, $\sigma^2_{\beta p} = 1$, $\sigma^2_{\pi p} = 0.1$, $\sigma^2_{\sigma\mu p} = \log(4)$, $\sigma^2_{\sigma\alpha p} = \log(8)$, $\sigma^2_{\sigma\beta p} = \log(1.5)$ and $\sigma^2_{\sigma\gamma p} = \log(1.5)$. Following Roberts et al. (1997), step sizes have been tuned so that the acceptance rates are around 20%.

Acceptance Probabilities (Metropolis-Hastings ratio)

Updates for μ

Choose a μ_k for $k = 2, \ldots, q-1$ at random and sample μ'_k from proposal $q(\mu'_k|\mu_{k-1}, \mu_k, \mu_{k+1})$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu', \alpha, \beta, \pi)P(\mu'|\sigma^2_\mu)}{P(Y|\mu, \alpha, \beta, \pi)P(\mu|\sigma^2_\mu)} \times \frac{\min(\mu_k + \tau, \mu_{k+1}) - \max(\mu_k - \tau, \mu_{k-1})}{\min(\mu'_k + \tau, \mu_{k+1}) - \max(\mu'_k - \tau, \mu_{k-1})}\right]$$

where $\mu = (\mu_1, \ldots, \mu_k, , \ldots, \mu_{q-1})$, $\mu' = (\mu_1, \ldots, \mu'_k, , \ldots, \mu_{q-1})$ for $k = 1, \ldots, q-1$ and $\mu_1 = 0, \mu_0 = -\infty$, $\mu_q = \infty$.

Updates for α

Choose a α_r for $r = 1, \ldots, R$ at random and sample α'_r from proposal $q(\alpha'_r|\alpha_r)$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, \alpha', \beta, \gamma, \pi)P(\alpha'|\sigma^2_\alpha)}{P(Y|\mu, \alpha, \beta, \gamma, \pi)P(\alpha|\sigma^2_\alpha)}\right]$$

Where $\alpha = (\alpha_1, \ldots, \alpha_r, , \ldots, \alpha_R)$ and $\alpha' = (\alpha_1, \ldots, \alpha'_r, , \ldots, \alpha_R)$.

Updates for β

Choose a k' from $k' = 1, \ldots, q$ at random and sample $\beta'_{k'}$ from proposal $q(\beta'_{k'}|\beta_{k'})$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, \alpha, \beta', \gamma, \pi)P(\beta'|\sigma^2_\beta)}{P(Y|\mu, \alpha, \beta, \gamma, \pi)P(\beta|\sigma^2_\beta)}\right]$$

Where $\beta = (\beta_1, \ldots, \beta_{k'}, , \ldots, 0)$ and $\beta' = (\beta_1, \ldots, \beta'_{k'}, \ldots, 0)$.

Updates for γ

Choose a j from $j = 3, \ldots, p$ at random and sample γ'_j from proposal $q(\gamma'_j|\gamma_j)$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, \alpha, \beta, \gamma', \pi)P(\gamma'|\sigma^2_\gamma)}{P(Y|\mu, \alpha, \beta, \gamma, \pi)P(\gamma|\sigma^2_\gamma)}\right]$$

Where $\gamma = (0, 0, \ldots, \gamma_j, , \ldots, \gamma_p)$ and $\gamma' = (0, 0, \ldots, \gamma'_j, , \ldots, \gamma'_p)$.

Updates for σ^2_μ, σ^2_α, σ^2_β and σ^2_γ

Given σ^2_μ, sample from σ'^2_μ proposal $q(\sigma'^2_\mu|\sigma^2_\mu)$ and accept with probability

$$r = \min\left[1, \frac{P(\beta|\sigma'^2_\mu)P(\sigma'^2_\mu)}{P(\beta|\sigma^2_\mu)P(\sigma^2_\mu)} \times \frac{\sigma^2_\mu}{\sigma'^2_\mu}\right]$$

Similarly for σ^2_α, σ^2_β and σ^2_γ.

Updates for π

Given π sample π' from $q(\pi'|\pi)$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, \alpha, \beta, \pi')P(\pi')}{P(Y|\mu, \alpha, \beta, \pi)P(\pi)} \times \frac{w'(1-w')}{w(1-w)}\right]$$

where $\pi = (\pi_1, \ldots, \pi_{r1}, \ldots, \pi_{r2}, \ldots, \pi_R)$, $\pi' = (\pi_1, \ldots, \pi'_{r1}, \ldots, \pi'_{r2}, \ldots, \pi_R)$, and $w = \pi_{r1}/(\pi_{r1} + \pi_{r2})$, $w' = \pi'_{r1}/(\pi'_{r1} + \pi'_{r2})$.

Notice that in the case of α, β and γ the proposal density $q(.) \sim \text{Normal}(.)$ is symmetric and thus cancels out from the MH ratio. Updates for σ^2_μ, σ^2_α, σ^2_β, σ^2_γ and π involve transformations so that a Jacobian included. The proposal for μ is not symmetric and thus it can not be dropped from the MH ratio.

Chapter 7

Bayesian Non-Parametric models

7.1 Introduction

This chapter develops Bayesian Non-Parametric (BNP) models for model-based clustering in repeated ordinal data. A more detailed discussion and applications could be found at Mitra & Müller (2015), Hjort et al. (2010) whereas more mathematical treatments are given by Ghosh & Ramamoorthi (2003), Phadia (2013).

A parametric model is a model that belongs to a family of finite dimensional models, that is a model that has a finite-dimensional parameter space. Let f be a function with parameter θ, then a parametric model is any of the family $S = \{f_\theta : \theta \in \mathbb{R}^{\mathrm{d}}\}$, where d is the cardinality of θ. Given that the true f is unknown, misspecification of this function is always a danger when using parametric models.

In contrast to that, a non-parametric (NP) model is a model that has a infinite-dimensional parameter space. That is, the parameter of the model is a set that has infinite elements. For instance, the space of positive functions v on the real line $S = \{v(x) : v(.) > 0, x \in \mathbb{R}\}$ has infinite dimension.

Any non-parametric model can be formulated as a semi-parametric model by separating its infinite dimensional parameter into infinite and finite parts. Let θ be the infinite dimensional parameter , we can always

re-express it as $\theta = (\theta_1, \theta_2)$, where $\theta_1 \in \mathbb{R}^{\mathrm{d}}$, and $\theta_2 \in S$ an infinite dimensional set. Non-Parametric Frequentist inference usually treats infinite-dimensional parameters as nuisance and leaves them unspecified, estimating only the finite-dimensional parameters. For instance, the well-known proportional hazards model, (Cox 1972):

$$\lambda(t, x) = \lambda_0(t)\exp(x'\beta)$$

leaves the baseline hazard $\lambda_0(t)$ unspecified and focuses only on the estimation of β, whose dimension is equal to the number of covariates (x). Note that $\lambda_0(t)$ belongs to a space of functions such that $\lambda_0(.) > 0$ and is thus an infinite dimensional parameter.

In contrast to this, BNP models place a prior for the infinite dimensional parameter and therefore provide a full probabilistic description of the model. Putting priors on infinite dimensional objects is of course not a trivial task. The development of the Dirichlet Process prior however provided a practical solution (Ferguson 1973). Technically, a BNP model is a probability model on infinite dimensional probability spaces. More precisely, given a measurable space $(\Omega, \mathcal{X})$, where Ω is a set and $\mathcal{X}$ is a σ-algebra on Ω, a BNP model assigns a prior to the probability space $(\Omega, \mathcal{X}, P)$. Here P is a measure that satisfies the axioms of probability. No theoretical results will be proven in this chapter so we will not use the above technical terminology when referring to the BNP models.

In this thesis, we are interested in using BNP models to make inference on probability distributions, followed by classification, and thus a BNP prior could also be seen as a random probability measure, that is a probability measure on a collection of distribution functions (Müller et al. 2015). In practice, this random probability measure is centered at a given parametric family and thus provides more flexibility and a more robust inference against misspecification of the parametric family. This centering family is known as the centering measure and has the role of guiding "the posterior distribution when the data are sparse, but allows the posterior to adapt locally where the data are plentiful" (Jara et al. 2008).

7.2 Dirichlet Process

Dirichlet Process (DP) Definition

Given $M > 0$, a probability measure G_0 with support S and any measurable finite partition $\{B_1, B_2 \ldots, B_k\}$ of S, Ferguson (1973) showed that a Dirichlet Process $DP(M, G_0)$ is a random probability measure G such that

$$(G(B_1), G(B_2), \ldots, G(B_k)) \sim \text{Dirichlet}(MG_0(B_1), MG_0(B_2), \ldots, MG_0(B_k))$$

where

$$\begin{aligned} E[G] &= G_0 \\ \text{Var}[G] &= \frac{G_0(1-G_0)}{(M+1)} \end{aligned} \tag{7.1}$$

G_0 is known as the centering measure and M as the precision or total mass parameter. The product MG_0 is called base measure of the DP. Under mild conditions, G can weakly approximate any distribution with the same support as G_0. Notice that the number of partitions k is always discrete, that is G is discrete with probability one.

Importantly, Ferguson (1973) also showed that a DP is a conjugate prior with respect to iid sampling. Let $y_1, y_2, \ldots y_n | G \overset{iid}{\sim} G$ and $G \sim DP(M, G_0)$ then

$$G|y_1, y_2, \ldots y_n \sim DP(M+n,\ \frac{MG_0 + \sum_{i=1}^n \delta_{y_i}}{M+n}) \tag{7.2}$$

where δ_{y_i} is a Dirac unit probability mass at location y_i. In words, if we have iid observations and we assign a DP prior for its unknown distribution, the posterior for G is also a DP with an updated precision $M+n$ and centering measure $\frac{MG_0 + \sum_{i=1}^n \delta_{y_i}}{M+n}$. Using (7.1), its expected value and variance can also be obtained

$$\begin{aligned} E[G|y_1, y_2, \ldots y_n] &= \frac{MG_0 + \sum_{i=1}^n \delta_{y_i}}{M+n} = \frac{G_0'}{M+n} \\ \text{Var}[G|y_1, y_2, \ldots y_n] &= \frac{G_0(1-G_0)}{(M+1)} = \frac{G_0'(M+n-G_0')}{(M+n)(M+n+1)} \end{aligned}$$

That is, the posterior mean of G can be seen as a weighted average of the centering measure G_0 and the empirical distribution of the sample $1/n \sum_{i=1}^{n} \delta_{y_i}$. Note also that for large n

$$G|y_1, y_2, \dots y_n \propto G_0' = MG_0 + \sum_{i=1}^{n} \delta_{y_i}$$

the variance of the posterior of G is very small so that sampling from it can be approximated by sampling directly from the (non-scaled) updated centering measure G_0'. Figure 7.1 presents draws y from a DP using a standard logistic distribution as a centering measure and different values of M. Note that the empirical cumulative distribution function of y is a step-function due to the discrete nature of the DP, i.e. there are ties in the values of y_i. As precision M increases, the steps of the empirical cumulative distribution of y become smaller and the $DP(M, G_0)$ gets closer to the centering measure G_0.

Pólya Urn representation

Blackwell & MacQueen (1973) showed that G could be marginalized out from (7.2) and that the distribution of $y_1, y_2, \dots y_n$ could be directly expressed as coming from a Pólya Urn scheme,

$$P(y_1, \dots, y_n) = P(y_1) \prod_{i=2}^{n} P(y_i|y_1, \dots, y_{i-1})$$

where

$$\begin{aligned} y_1 &\sim G_0 \\ P(y_i|y_1, \dots, y_{i-1}) &= \frac{1}{M+i-1} \sum_{h=1}^{i-1} \delta_{y_h(y_i)} + \frac{M}{M+i-1} G_0(y_i),\ i \geq 2 \end{aligned} \tag{7.3}$$

This scheme postulates that after the first draw y_1, that comes from the centering measure G_0, draws y_i could either be equal to one of the previous ones $y_1 \dots y_{i-1}$, with probability proportional to $\frac{1}{M+i-1}$, or come from the

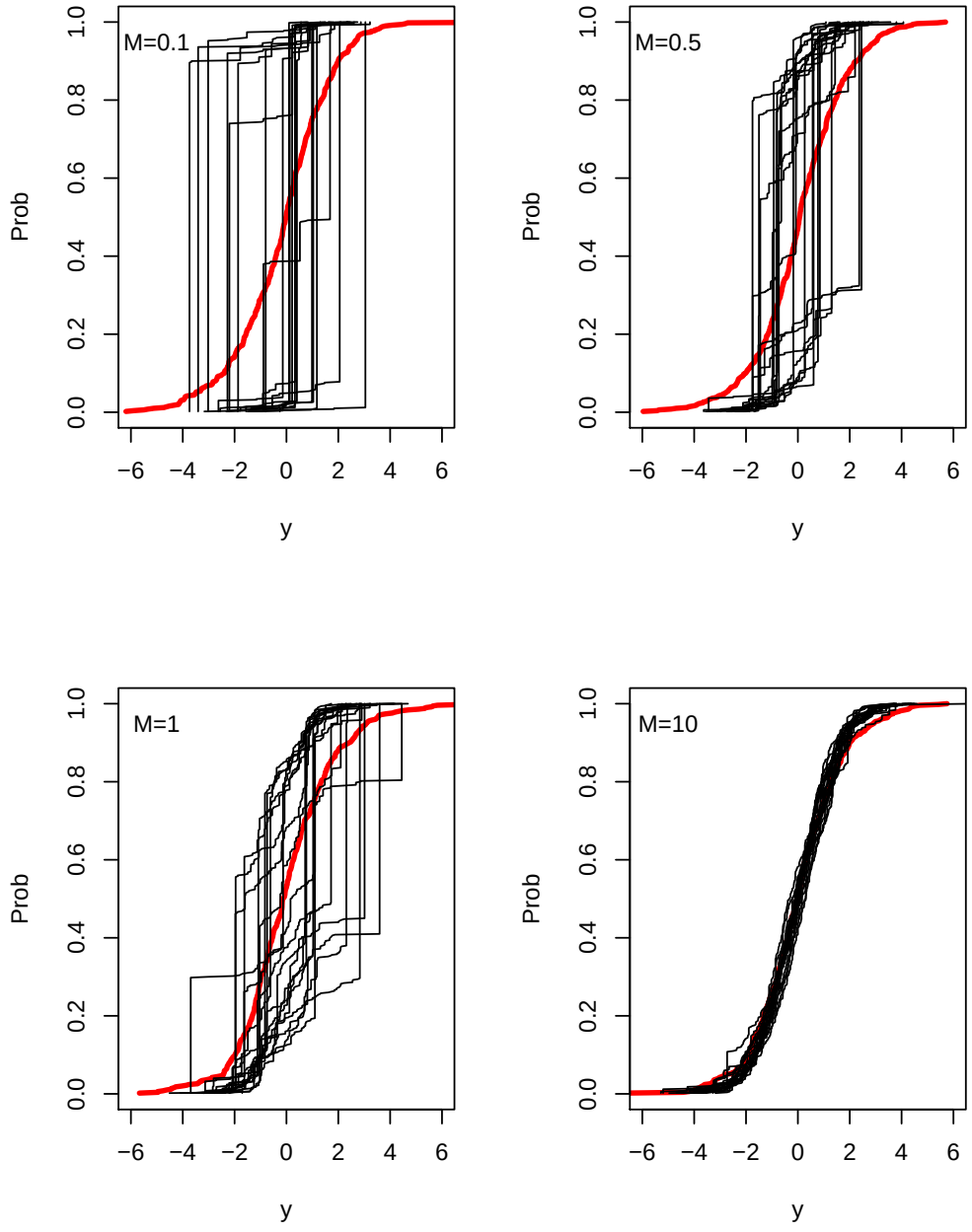


Figure 7.1: Empirical cumulative distribution function of draws from $y \sim DP(M, G_0)$ with centering measure $G_0 \sim$ Logistic(0, $\pi^2/3$), standard logistic distribution, and varying precision parameter, $M = 0.1, 0.5, 1, 10$. Graphs display 20 draws y_i of size 500 each. Thick red line in the middle shows $E[G] = G_0$.

centering measure G_0, with probability proportional to $\frac{M}{M+i-1}$. This representation of the DP is also known as Chinese restaurant process, where arriving customers can either join any of the already existing tables, occupied by some customers with probability proportional to the number of people already seated on the table, or sit at a fresh new table on its own.

Stick breaking construction

As a discrete random probability measure, Sethuraman (1994) showed that G could also be written as a infinite weighted sum of point masses $G(.) = \sum_h^\infty w_h \delta_{m_h}(.)$ where $w_1, w_2, \ldots$ are probability weights and $\delta_x(.)$ denotes the point mass at location x.

Let $w_h = v_h \prod_{\ell<h}(1 - v_\ell)$ with $v_h \overset{iid}{\sim}$ Beta$(1, M)$ and $m_h \overset{iid}{\sim} G_0$ then

$$G(.) = \sum_h^\infty w_h \delta_{m_h}(.) \tag{7.4}$$

defines a $DP(M, G_0)$ random probability measure. Figure 7.2 shows a graphical representation of this stick breaking construction of the DP. Locations m_h are sampled from G_0 while their corresponding weights w_h from the remaining of the stick using a Beta distribution $v_h \sim$ Beta$(1, M)$.

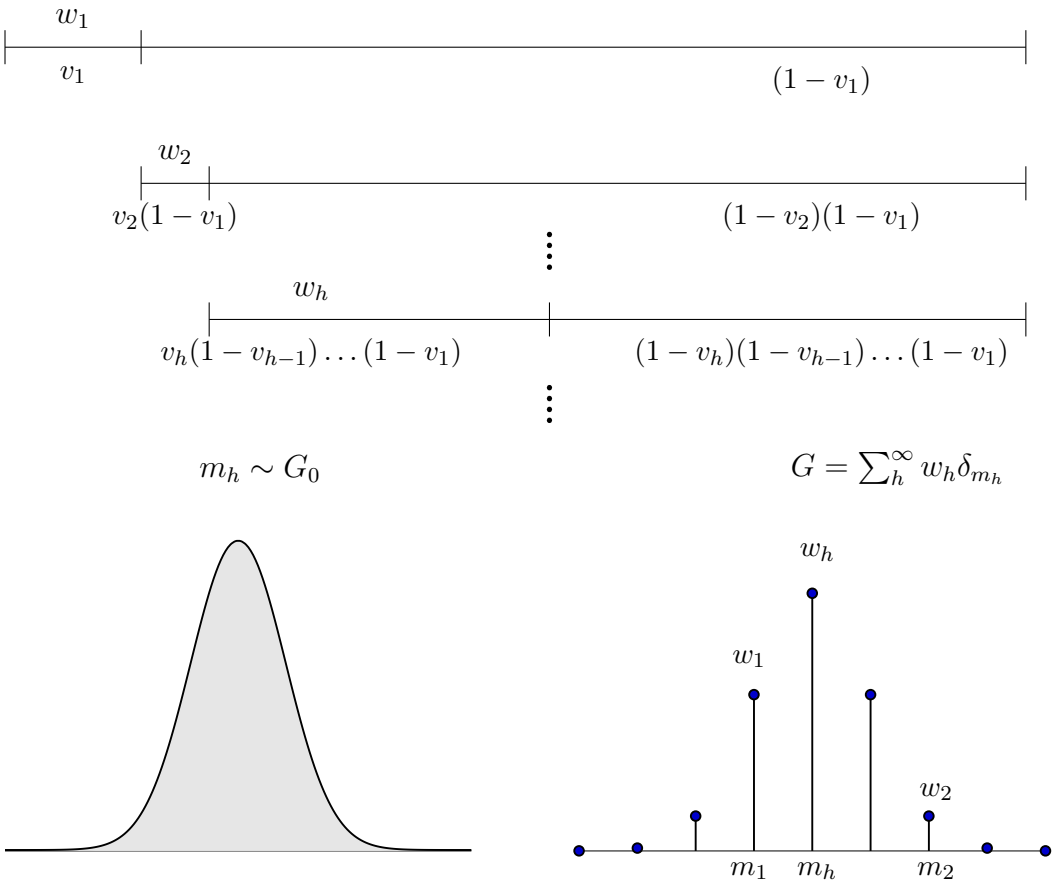


Figure 7.2: Stick breaking construction of the DP

7.3 Dirichlet Process Mixture

Dirichlet Process Mixture (DPM) Definition

Let Θ be a finite-dimensional parameter space, for each $\theta \in \Theta$, f_θ is a continuous pdf. Given G defined on Θ, a mixture of f_θ with respect to G has pdf

$$f_G(y) = \int f_\theta(y) dG(\theta) \tag{7.5}$$

f_θ is known as the mixture's kernel and G as the mixing probability (Ferguson 1983). The choice of the appropiate kernel depends on the underlying sample space. In this case, the hierarchical representation of the joint distribution of the DPM and the data $i = 1, \ldots, n$ is

$$\begin{aligned} y_i|\theta_i &\overset{ind}{\sim} f_{\theta_i} \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim DP(M, G_0) \end{aligned} \tag{7.6}$$

or equivalently

$$\begin{aligned} &y_i|f_G \overset{iid}{\sim} f_G \text{ where } f_G = \int f_\theta(y) dG(\theta) \\ &G \sim DP(M, G_0) \end{aligned}$$

DPM is also conjugate prior, similarly to 7.2 a DPM is conjugate with respect to iid sampling. Let $i = 1, \ldots, n$ and

$$\begin{aligned} y_i|\theta_i &\overset{ind}{\sim} f_{\theta_i} \\ \theta_i|G &\overset{iid}{\sim} G \implies G|y \sim \int DP(MG_0 + \sum_{i=1}^{n} \delta_{\theta_i}) dP(\theta|y) \\ G &\sim DP(MG_0) \end{aligned} \tag{7.7}$$

That is, if an unknown pdf f with finite-dimensional parameter, θ is given a DPM prior, its posterior distribution is also a DPM with an updated base measure (original MG_0 plus point masses located at each observation). Additionally, for large n, $G \sim DP(MG_0 + \sum_{i=1}^{n} \delta_{\theta_i})$ can be approximated by $G \propto MG_0 + \sum_{i=1}^{n} \delta_{\theta_i}$.

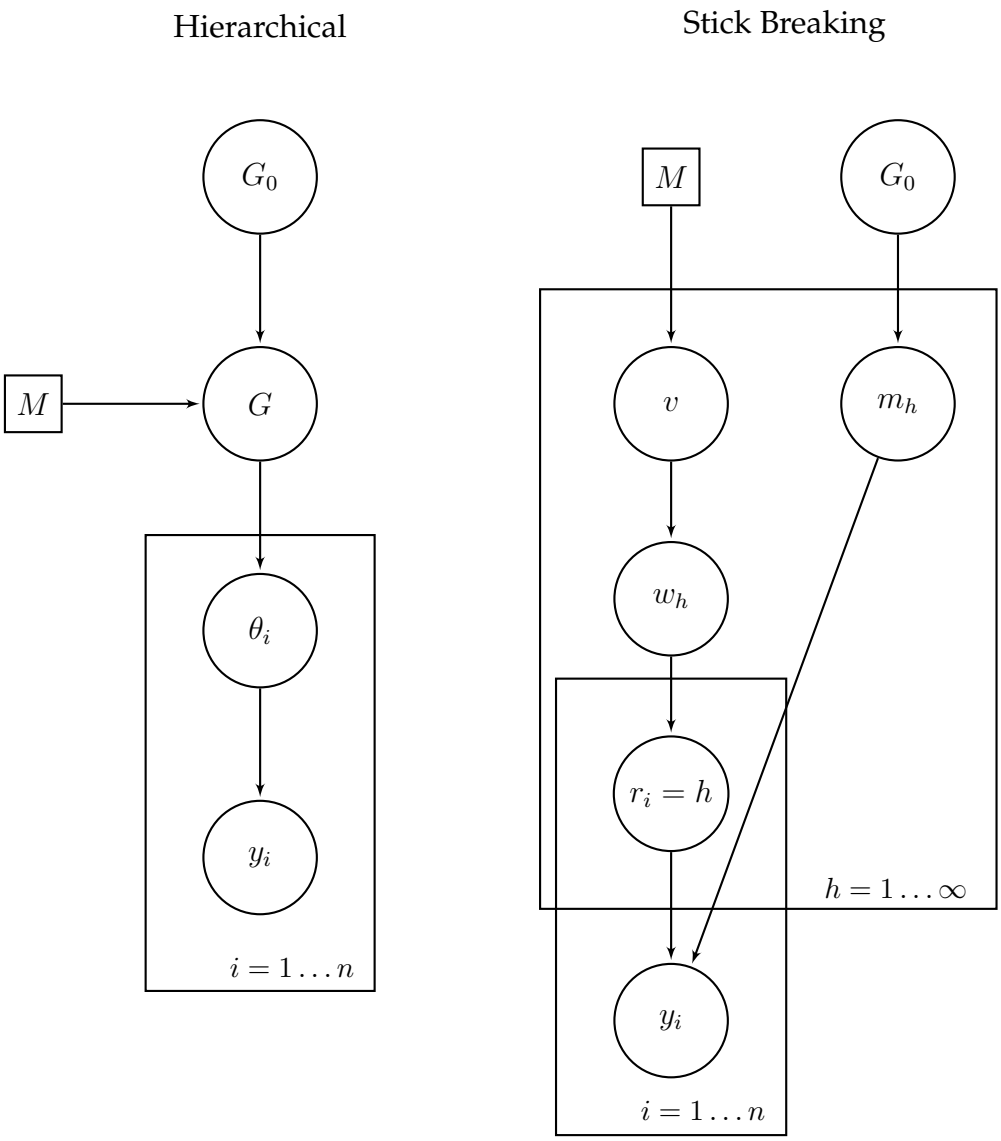


Figure 7.3: Alternative graphical representations of the DPM

Alternative representations of a DPM

Firstly, given the finite nature of $\theta = (\theta_1, \theta_2, \dots, \theta_n)$, it will have at most n elements and there may possibly be ties. Let θ_j^* the unique values among $\theta_1, \theta_2, \dots, \theta_n$ and indicator variables $s_i = 1, \dots, K$ such that $\theta_i = \theta_{s_i}^*$, then

(7.6) could also be written as

$$\begin{aligned} y_i|s_i, \theta_j^* &\sim f_{\theta_{s_i}^*} \\ \theta_j^* &\sim G_0 \end{aligned}$$

where $P(s_1, s_2, \ldots, s_n) = \frac{\Gamma(M)}{\Gamma(M+n)} M^k \prod_{j=1}^{k} \Gamma(n_j)$ is the joint probability of cluster memberships and $n_j = \sum_i I(s_i = j)$ number of elements of cluster j.

Secondly, and similarly to (7.3), G can be marginalized from the DPM to get its Pólya Urn representation

$$\begin{aligned} y_i|\theta_i &\sim f_{\theta_i} \\ (\theta_1, \theta_2, \ldots, \theta_n) &\sim P(\theta_1) \prod_{i=2}^{n} P(\theta_i|\theta_1, \ldots, \theta_{i-1}) \end{aligned}$$

where $P(\theta_i|\theta_1, \ldots, \theta_{i-1}) \propto \sum_{j=1}^{K_n} \eta_{n_j} \delta_{\theta_j^*} + MG_0$

7.3.1 Finite Dirichlet Process Mixture (DPM$_H$)

Ishwaran & Zarepour (2000), Ishwaran & James (2001, 2002) proposed to approximate a DPM with a finite sum in (7.4). A finite DPM is thus a DPM truncated to an upper bound H, where truncation is done by setting $v_H = 1$

$$\begin{aligned} G(.) =& \sum_{h=1}^{H} w_h \delta_{m_h}(y) \\ w_h =& v_h \prod_{l<h} (1 - v_l) \text{ and } v_h \overset{iid}{\sim} \text{Beta}(1, M) \; ; h = 1, \ldots, H-1; v_H = 1 \\ m_h \overset{iid}{\sim}& G_0; h = 1, \ldots, H \end{aligned} \tag{7.8}$$

Hierarchical representation

$$y_i|f_G \sim f_G \qquad i = 1, \ldots, n \qquad \Longleftrightarrow \qquad \begin{array}{l} y_i|r_i = h \sim f_{m_h} \\ P(r_i = h) = w_h \end{array} \qquad i = 1, \ldots, n; h = 1, \ldots, H$$

Where $f_G = \sum_{h=1}^{H} w_h \delta_{m_h}(y)$. Importantly, Ishwaran & Zarepour (2000) also showed that H can be set to achieve a desired level of accuracy in the approximation of G. They showed that the expected value of the tail probability is

$$E\left[\sum_{h=H+1}^{\infty} w_h\right] = \left(\frac{M}{M+1}\right)^H$$

For instance, when $M = 1$ and $H = 20$, the expected value of this left-out probability is $9.536 \times 10^{-7} < 1 \times 10^{-6}$. We will use these values of M and H when fitting a finite DPM for repeated ordinal (Sections 7.6 and 7.7).

When G_0 is a conjugate prior for the mixture's kernel f_m, a Gibbs sampler can be use to simulate the posterior of a DPM$_H$. Müller et al. (2015) proposed the following general algorithm to sample from $p(r, v, m|y)$:

Algorithm 1: MCMC sampler for DPM$_H$

1. Cluster memberships r_i:
   $r_i \sim$ Categorical$(P(r_i = 1|v, m, y_i), \ldots, P(r_i = H|v, m, y_i)), i = 1, \ldots n$
   $P(r_i = h|v, m, y_i) \propto w_h f_{m_h}(y_i),\ h = 1, \ldots H$

2. Weights w_h:
   $v_h \sim$ Beta$(1 + A_h, M + B_h), h = 1, \ldots, H-1, v_H = 1$
   $w_h = v_h \prod_{l<h}(1 - v_l)$, $h = 1, \ldots, H$
   $A_h = \sum_i I(r_i = h)$ observations assigned to cluster h, and
   $B_h = \sum_i I(r_i > h)$ observations not yet assigned to any cluster.

3. Locations m_h:
   $m_h \sim P(m_h|y, r) \propto G_0(m_h) \prod_{i \in S_h} f_{m_h}(y_i)$, h=1,$\ldots H$
   $S_h = \{i : r_i = h\}$ denotes the set of observations with $r_i = h$

Notice that some S_h may be empty, that is no observations i have been assigned to cluster h. Importantly, this algorithm is very general and can also be used when G_0 and f_m are not conjugate by replacing steps 1 and 3 with appropriate Metropolis-Hastings transitions.

7.4 Post-hoc clustering of clusters

As seen before in (7.3), a DP is almost surely discrete which implies a positive probability of ties among its realizations. Further to that, Korwar & Hollander (1973) showed an analytic expression for the expected number of distinct values K,

$$\begin{aligned} E[K] = &\sum_{i=1}^{n} \frac{M}{M+i-1} \\ \approx &M \log(n) \qquad\qquad (\text{as } n \implies \infty) \end{aligned} \tag{7.9}$$

In other words, the expected number of distinct values K, active point masses or non-empty clusters, in a sample that follows a DP is asymptotically smaller than n ($M \log(n) << n$). A very important implication for our model-based clustering purposes is that K grows (slowly) with n and this is not appropriate if we believe that there is a true model whose dimension, number of latent components, does not depend on the sample size. Importantly however, a small number of the ties repeat often and large number repeat rarely so that there are a small number of big-size groups and a big number of groups that are singletons or have a very small size.

For instance, Figure 7.4 displays the posterior distribution of K for a Poisson mixture with three components estimated using a finite DPM. The mixture has equal proportions and rates $\lambda = (1, 4, 8)$, and a base measure $G_0 \sim$ Gamma(1,0.05) which is conjugate to the Poisson likelihood. Setting $H = 100$, $M = 1$ and taking a sample of $n = 500$ observations, we estimate the DPM$_H$ using Algorithm 1. For this sample, it can be seen that

the most likely values for K are 4 to 6 (with $E[K] \approx 6.21$) in the upper panel. In contrast to that, when we focus on groups that not so small, for instance groups have at least 5% of the sample size (25 observations) in the bottom panel, the distribution for K changes markedly. Now, the number "non-neglible" clusters is most likely to be 3 or 4 (the former being the true number of mixture components).

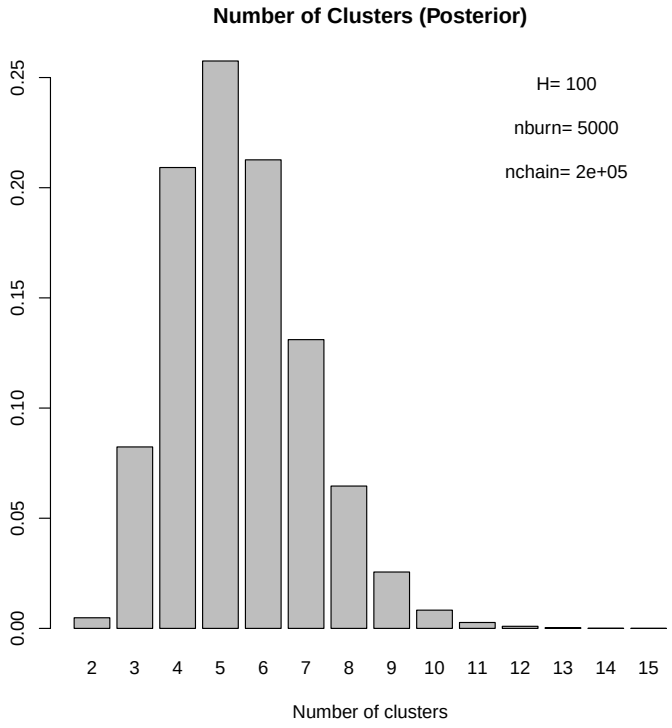
Although different, the point masses of a DPM are likely to be located around similar places if the data is truly being generated by a finite mixture. We can therefore post-process all these locations from the DPM to see if any pattern emerges. We called this procedure post-hoc clustering of clusters, i.e. clustering the locations m_h estimated by the DPM. Fraley & Raftery (2002) pointed out that model based clustering can also be useful to deal with such a "meta-problem". For instance, given $m_h \in \mathbb{R}$, a mixture of normals could be used for this aim.

In contrast to that, we use here two non-model based approaches: k-means clustering of the posterior distribution of all the locations of the DPM model and hierarchical clustering using the estimated distance matrix (complement of the co-clustering matrix). Moreover, both approaches also allow to deal with label switching. An alternative approach is presented by Dahl (2006), who uses the MCMC draw that minimises the squared distance to the co-clustering matrix, as the optimal partition.

K-means clustering of all DPM locations

One way to cluster the locations of the DPM would be to use a partition optimization method such as the k-means algorithm Lewis et al. (2003). We begin by plotting the distribution of the locations over the Markov chain. Doing so, will reveal any pattern and multimodality in this distribution. A histogram and kernel estimator can be useful to estimate the number of potential modes.

Secondly, given that the labels and its number (K) vary over MCMC iterations we apply the k-means algorithm to the locations over all the



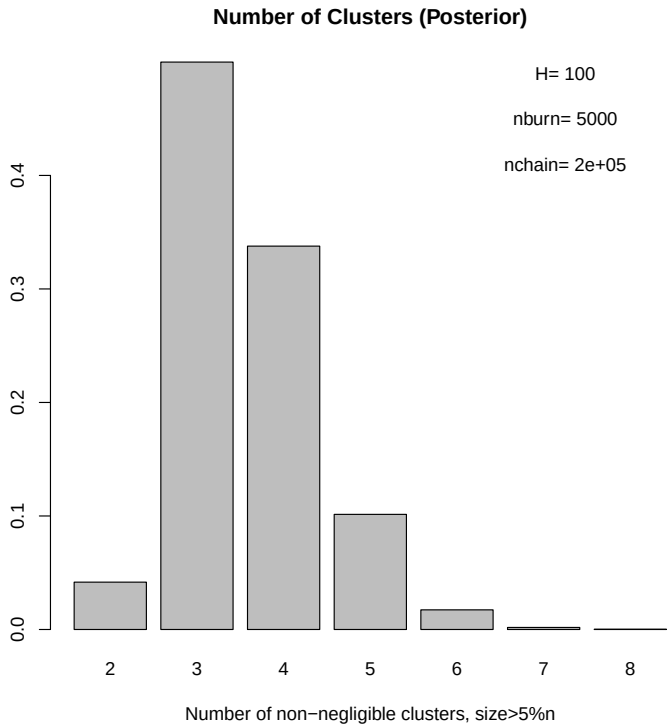


Figure 7.4: Posterior distribution of K (number of clusters) for a simulated Poisson mixture with three components with equal proportions and rates $\lambda = (1, 4, 8)$ and $n = 500$. Mixture is estimated using a finite DPM with $G_0 \sim$ Gamma(1,0.05), $H = 100$, and $M = 1$.

Markov chain. That is, if S is the number of MCMC draws and we have n responses, we use the k-means algorithm to cluster $S \times n$ locations into the number of modes obtained above. Due to the possibility of observations being allocated to different groups over MCMC, we further allocate each subject to the group where it appeared most often.

Let $c = 1, \ldots, C$ be the number of modes of the distribution of the DPM locations over the MCMC iterations $s = 1, \ldots, S$, then the probability that individual i comes from mode c, z_{ic} is

$$\hat{z}_{ic} = \frac{\sum_{s=1}^{S} I\{l_i = c\}}{S}$$

Note that, by clustering the locations into C groups over the MCMC chain, and not at each iteration, there is no label switching. After that, allocation l_i of individual i to mode c can based on a highest a posteriori probability criterion:

$$\hat{l}_i = \underset{c \in 1,\ldots,C}{\operatorname{argmax}} \, \hat{z}_{ic} \ , \ i = 1 \ldots n$$

In this dissertation we use this procedure to carry out the post-hoc clustering of clusters. Examples of this procedure are presented in Figures 7.6 and 7.9.

Hierarchical clustering based on the distance matrix

Let G be the DPM, we define the co-clustering probability of observations i and j as:

$$p_{ij} = \frac{\sum_{s=1}^{S} I(r_i = h, r_j = h|Y, G)}{S} \equiv 1 - d_{ij} \tag{7.10}$$

where: S is the number of MCMC draws and d_{ij} the distance between observations i and j. The co-clustering probability matrix P, also known as average incidence matrix or pairwise probability matrix, is an n by n

matrix that gathers all the p_{ij}. Notice that p_{ij} is invariant to label switching because it compares the allocation of both observations to the same cluster h, regardless of the naming of component h at that iteration. Futhermore, it is also invariant to the number of non-empty components at iteration s.

Distances are then calculated as the complement of co-clustering probabilities $d_{ij} = 1 - p_{ij}$ and the cluster memberships for all individuals are finally obtained using hierarchical clustering (Kaufman & Rousseeuw 1990) with these distances. Molitor et al. (2010) and Si et al. (2014) have used this postprocessing procedure to cluster survey and RNA expression cross-sectional data, respectively. In this dissertation we use this procedure mainly for visualisation purposes, dendograms in Figures 7.7 and 7.10, after the k-means post-hoc clustering of all locations.

In the following sections, we present two examples to show how BNP models work in practice for repeated ordinal data. The former is a simulation and the latter is a case study where a BNP model is fitted to the life satisfaction in New Zealand dataset, already analysed in Chapter 5.

7.5 A DPM model for repeated ordinal data

We now use a finite Dirichlet Process Mixture to estimate a model for repeated ordinal data. This model is similar to the one in Chapter 5 (Parameter dependent models) where the occasion effects follow a random walk, ie $\beta_j \sim \text{Normal}(\beta_{j-1}, \sigma^2)$.

Using the same notation we used previously, we model the ordinal response y_{ij} as

$$
\begin{aligned}
&y_{ij} \mid (\mu_k), \alpha_i, \beta_j \sim \text{ Categorical}_q(\theta_{ij.}), \\
&\qquad \theta_{ijk} = \frac{1}{1+e^{-(\mu_k-\alpha_i-\beta_j)}} - \frac{1}{1+e^{-(\mu_{k-1}-\alpha_i-\beta_j)}} \, , \sum_{k=1}^{q} \theta_{ijk} = 1 \\
&\mu_k \mid \sigma^2_\mu \sim \text{Normal}(0, \sigma^2_\mu)\ \text{I}[\mu_k > \mu_{k-1}]; \quad k = 1, \ldots, q \\
&\beta_j \mid \sigma^2_\beta \sim \text{Normal}(\beta_{j-1}, \sigma^2_\beta); \quad j = 1, \ldots, p;\ \beta_1 = 0 \\
&\alpha_i = m_h \mid r_i = h \sim N(0, \sigma^2_\alpha); \quad i = 1, \ldots, n; \quad h = 1, \ldots, H \\
&w_h = P(r_i = h) = v_h \prod_{l<h} (1 - v_l) \text{ and } v_h \overset{iid}{\sim} \text{Beta}(1, M);\ h = 1, \ldots, H-1; v_H = 1 \\
&\sigma^2_\mu \sim \text{Inverse Gamma}(a_\mu, b_\mu) \\
&\sigma^2_\alpha \sim \text{Inverse Gamma}(a_\alpha, b_\alpha) \\
&\sigma^2_\beta \sim \text{Inverse Gamma}(a_\beta, b_\beta)
\end{aligned}
\tag{7.11}
$$

with hyperparameters: $M = 1, a_\mu = a_\alpha = a_\beta = 3$ and $b_\mu = b_\alpha = b_\beta = 5$.

Notice that this model also could be seen as a Dependent Dirichlet Process (DPP) (MacEachern 1999, De Iorio et al. 2004) where the dependence is introduced only through the locations and with a logit link for the latent variable that generates the ordinal response. Also, Bao & Hanson (2015), DeYoreo & Kottas (2017) presented DPP models for multivariate ordinal data in cross-sectional settings. These models include covariates exploiting the normal latent variable representation of ordinal data. We choose the above formulation to facilitate comparsion with previous chapters.

7.5.1 Construction of the MCMC chain

Given the logit link for ordinal data, there are no conjugate priors for μ and β. In contrast to that, and given that the DP is a conjugate prior, the posterior distribution is also a DP with updated parameters (see 7.2). We therefore use Metropolis-Hasting (MH) transitions for μ and β and Gibbs samplers for the DP locations (m_h) and weights (w_h) and the cluster memberships (r_i). Variances are also simulated using a Gibbs sampler.

Updates for μ (MH)

Choose a μ_k for $k = 2, \ldots, q-1$ at random, sample μ'_k from proposal $q(\mu'_k|\mu_{k-1}, \mu_k, \mu_{k+1}) = U[\max(\mu_k - \tau, \mu_{k-1}), \min(\mu_k + \tau, \mu_{k+1})]$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu', \alpha, \beta)P(\mu'|\sigma^2_\mu)}{P(Y|\mu, \alpha, \beta)P(\mu|\sigma^2_\mu)} \times \frac{\min(\mu_k + \tau, \mu_{k+1}) - \max(\mu_k - \tau, \mu_{k-1})}{\min(\mu'_k + \tau, \mu_{k+1}) - \max(\mu'_k - \tau, \mu_{k-1})}\right]$$

where $\tau = 0.1, \mu = (\mu_1, \ldots, \mu_k, , \ldots, \mu_{q-1})$, $\mu' = (\mu_1, \ldots, \mu'_k, , \ldots, \mu_{q-1})$ for $k = 1, \ldots, q-1$ and $\mu_1 = 0, \mu_0 = -\infty$, $\mu_q = \infty$. Note that this is the same scheme used for μ in previous chapters.

Updates for β (MH)

Sample β' from proposal $q(\beta'|\beta) = \text{Normal}_{p-1}(\beta, \sigma^2_{\beta p} I)$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, \alpha, \beta')P(\beta'|\sigma^2_\beta)}{P(Y|\mu, \alpha, \beta)P(\beta|\sigma^2_\beta)}\right]$$

Where $\beta = (0, \ldots, \beta_2, \ldots, \beta_p)$, $\beta' = (0, \beta'_2, \ldots, \beta'_p), \sigma^2_{\beta p} = 0.1$ and I is an identity matrix with rank $p-1$. Note that the only difference of this scheme and the ones used in previous chapters for β is the multivariate proposal.

Updates for m (MH)

Sample m' from proposal $q(m'|m) = \text{Normal}_H(m, \sigma^2_{mp} I)$ and accept with probability

$$r = \min\left[1, \frac{P(Y|\mu, m, \beta')P(m'|\sigma^2_m)}{P(Y|\mu, m', \beta)P(\beta|\sigma^2_m)}\right]$$

Where $m = (m_1, \ldots, m_H)$, $m' = (m'_1, \ldots, m'_H)$ and $\sigma^2_{mp} = 0.1$. I is an identity matrix with rank H. Note that the only difference of this scheme and the ones used previous chapters for α is the multivariate proposal.

Updates for w_h and r_i

We sample w_h and r_i using the following:

1. Cluster memberships r_i:
   $r_i \sim \text{Categorical}_H(P(r_i = 1|.), \ldots, P(r_i = H|.)), i = 1, \ldots n$
   $P(r_i = h|.) \propto w_h \prod_{j=1}^{p} \text{Categorical}_q(y_{ij}|\mu, \beta_j, m_h),\ h = 1, \ldots H$

2. Weights v_h:
   Letting $A_h = \sum_i I(r_i = h)$ and $B_h = \sum_i I(r_i > h)$ then
   $v_h \sim \text{Beta}(1 + A_h, M + B_h), h = 1, \ldots, H-1, v_H = 1$
   $w_h = v_h \prod_{l<h}(1 - v_l)$, $h = 1, \ldots, H$

The derivation of the full conditionals could be found in Appendix 1 at the end of the chapter.

Updates for σ^2_α, σ^2_β and σ^2_μ (Gibbs)

Given that the prior for the $\sigma^2_\alpha \sim$ Inverse Gamma(a_α, b_α) is conjugate to the distribution of $\alpha_i \mid r_i = h \sim N(0, \sigma^2_\alpha)$, the posterior for σ^2_α proportional to

$$\sigma^2_\alpha \mid \alpha_1, \ldots \alpha_n \propto \text{Inverse Gamma}(a_\alpha + n/2, b_\alpha + \sum_{i=1}^{n} \alpha_i/2)$$

with hyperparameters: $a_\alpha = 3$ and $b_\alpha = 5$. Updates for σ^2_β and σ^2_μ are obtained in a similar fashion.

7.6 Simulations

We now proceed to validate the model in (7.11) using simulated ordinal responses from a four component mixture with parameters:

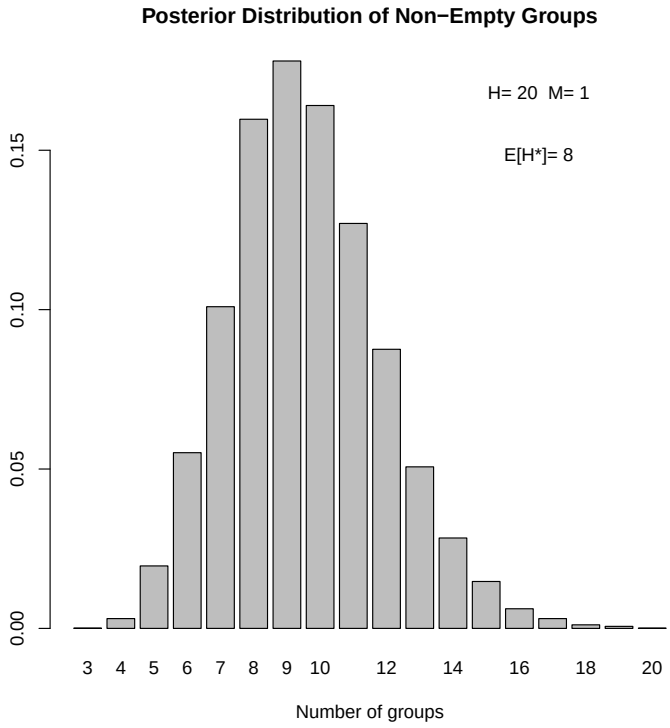
$$\begin{aligned}
n =& 3000,\ p = 5,\ q = 5 \\
\mu =& (0, 1.0, 1.8, 2.8) \\
\alpha =& (-2.5, -0.4, 1.44.2), \\
\beta =& (0, -2.2, -1.0, 0.9, 0.4),\ \sigma^2_\beta = 1.1 \\
\pi =& (0.14, 0.20, 0.24, 0.42).
\end{aligned}$$

The above settings imply a synthetic dataset with similar charactetistics to the NZAVS dataset to be used later in the case study. In terms of the parameters for the finite DPM, we use a truncation value of $H = 10$ and the hyperparameters detailed in (7.11). Discarding the initial 2500 draws as burn-in we used for inference 125000 MCMC draws (25000 thinned by 5). Figures 7.5 and 7.6 display the results.

Firstly, Figure 7.5 shows the posterior distribution of the non-empty locations of the DPM (K) given the data. Importantly, the expected number of non-empty groups could be easily computed using (7.9)

$$E[K] = \log(n) \times M = \log(3000) \times (1) = 8.0064$$

This is what we observe in the top panel, as the mode of the posterior is 9. However, due to the use of the DPM prior, this posterior distribution has a lot of singleton and small groups and a few bigger groups. For instance, in the bottom panel the distribution of groups with more than 15 observations (0.5% of the sample size 3000) has a mode of 6. Given that $E[K]$ grows with sample size, setting a threshold for what constitutes a small group in real-life applications would be arbitrary.



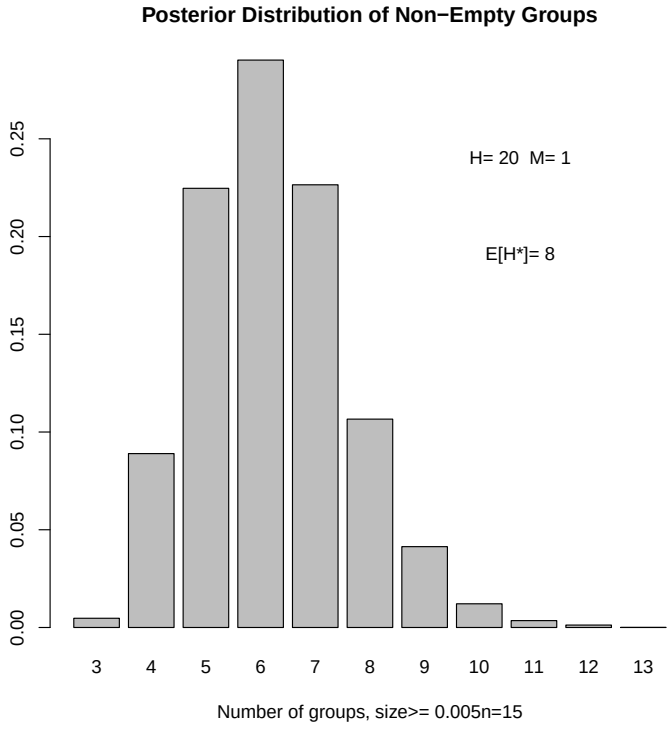


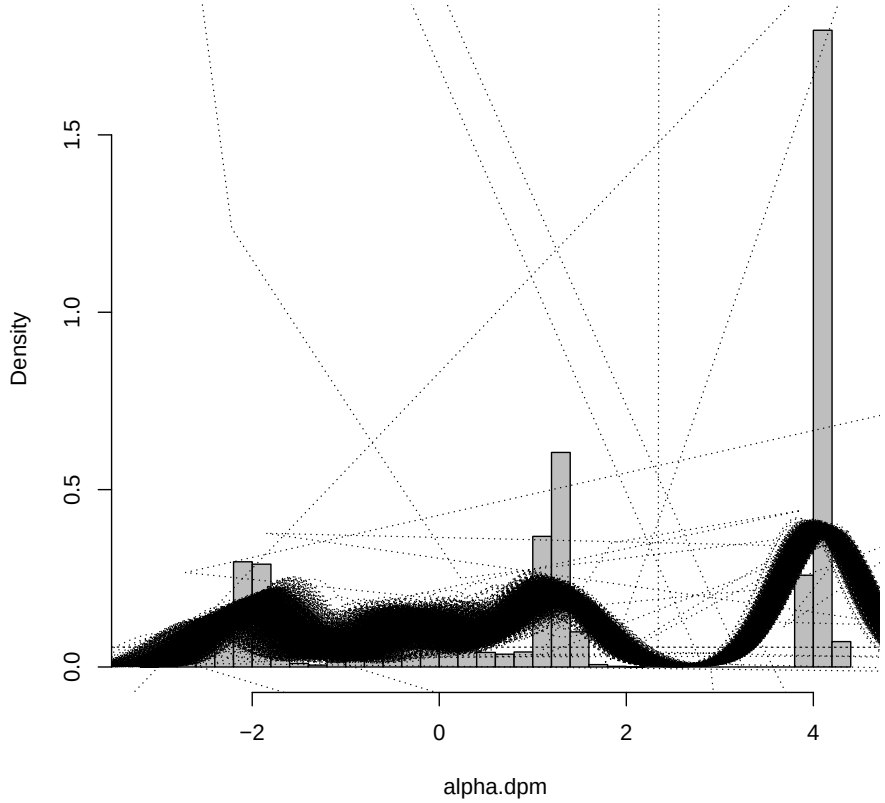
Figure 7.5: Posterior distribution for K for a DPM for the simulated ordinal data ($n = 3000, p = 5, q = 5$). True model is a four-component mixture with $\mu = (0, 1, 1.8, 2.8), \alpha = (-2.5, -0.4, 1.44.2), \beta = (0, -2.2, -1.0, 0.9, 0.4)$, and $\pi = (0.14, 0.20, 0.24, 0.42)$.

In contrast to that, the top panel of Figure 7.6 shows a histogram of the posterior distribution of the DPM locations (α_i). This figure also shows the distribution of locations for each MCMC iteration (black dotted lines) and its expected value over all MCMC iterations (red line). Here we can see that although a different number of point masses might be used on each MCMC draw (ranging from $h =, \ldots, H$), they tend to gather around a few common places and form a multimodal distribution with four modes around the true values (blue vertical dash lines) and modal densities proportional to the true mixture proportions. We therefore, set $C = 4$ the number of modes of the posterior distribution of the DPM locations. Notice finally that, there are three clusters of locations that are relatively closer to each other and this will be reflected in the uncertainty of their associatted classification.

In addition to that, the lower panel of Figure 7.6 shows a heatmaps of the co-clustering probabilities p_{ij} from the DPM model ordered by $C = 4$ clustering of clusters using the k-means clustering of all DPM locations (Section 7.4). The ordered heatmap of co-clustering probabilities shows that the classification probalities into any of the four clusters of locations are pretty good, except for the smallest group, where it is around 40%, which corresponds to the cluster of locations near zero in the top panel. On the other hand, classification probabilities in the biggest cluster of locations are the highest since it is relatively further away from the others (cluster around 4).

An alternative way of visualising the clustering of locations is shown Figure 7.7. This figure displays downwards and unrooted dendograms constructed using hierarchical clustering of the distance probability matrix as detailed in Section 7.4. We can see that, the four clusters of locations, shown in colours, are also very evident. Interestingly, the unrooted dendongram is the most revealing as it clearly shows that there is one cluster of locations that is further away from the other.

In sum, we can say that for this simulated dataset, the finite DPM



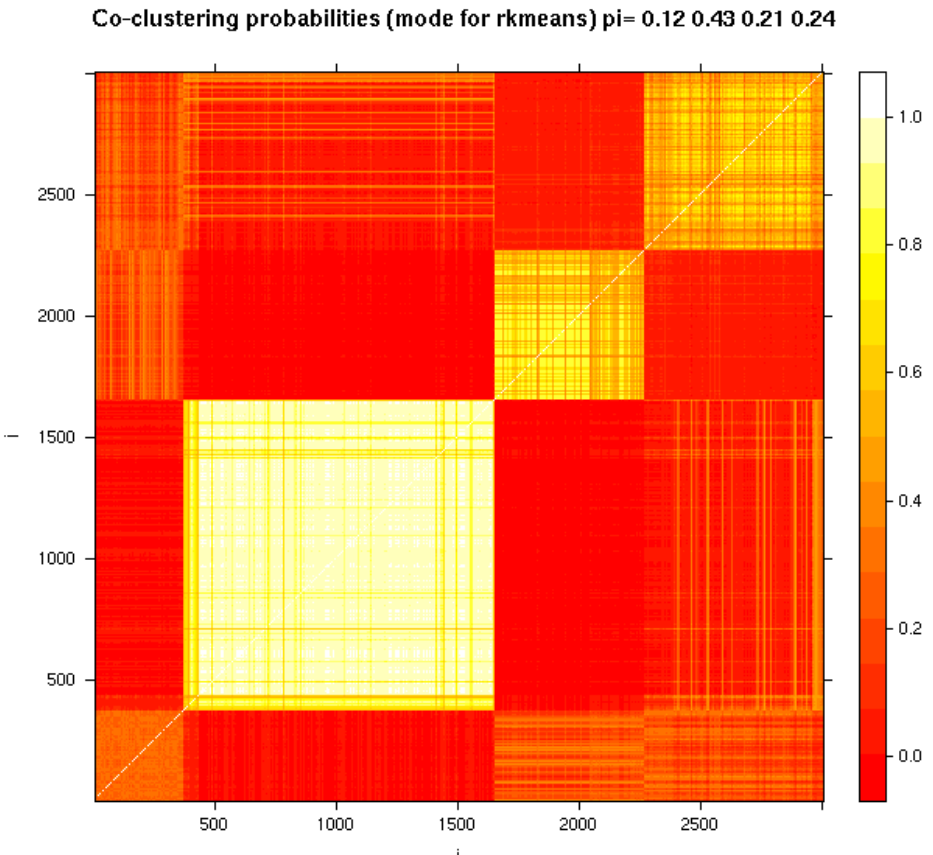


Figure 7.6: Posterior distribution of the locations (top) and ordered co-clustering probabilities (bottom) for the simulated ordinal data. True model is a four-component mixture with $\pi = (0.14, 0.20, 0.24, 0.42)$.

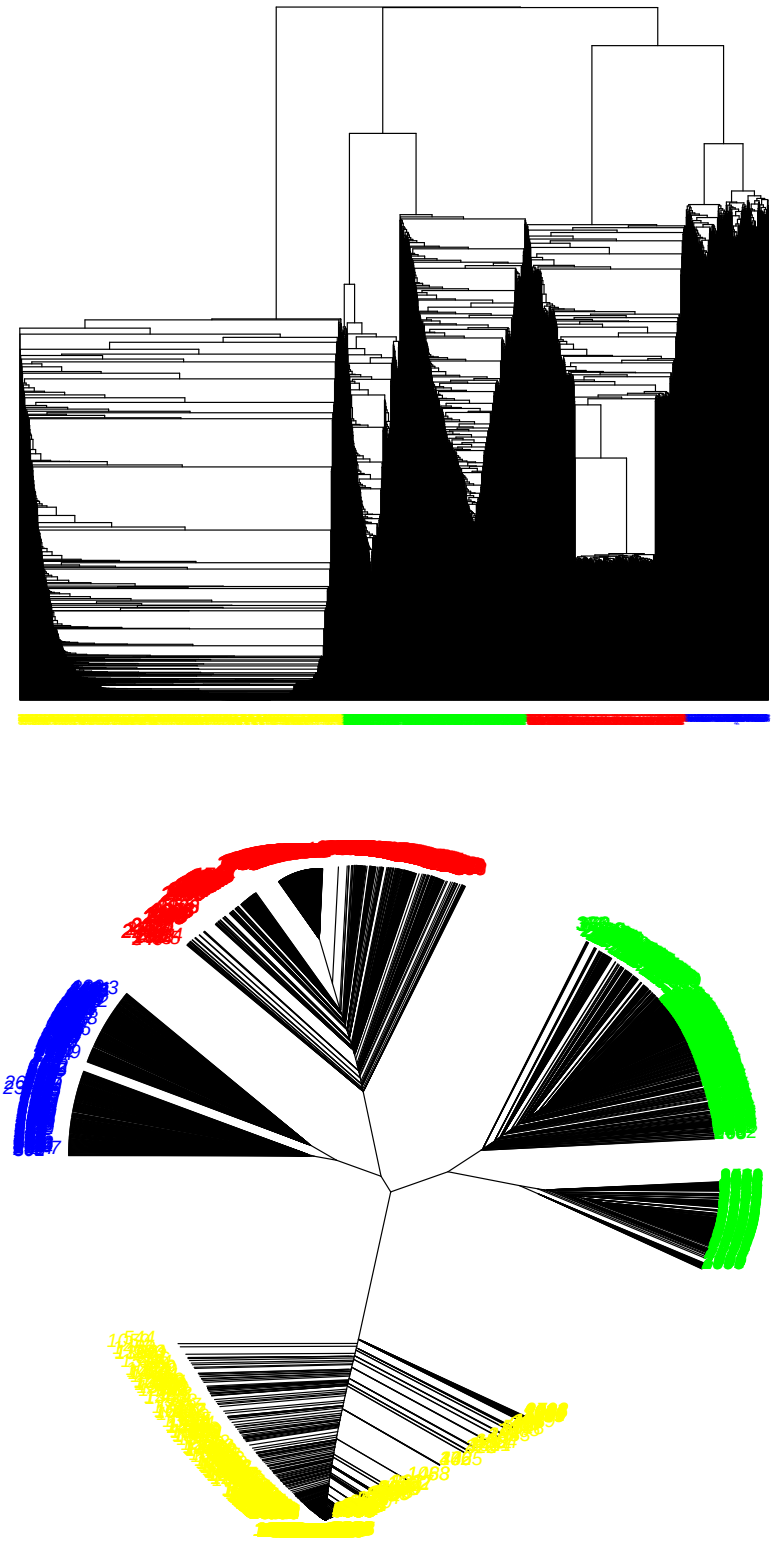


Figure 7.7: Dendograms for the simulated ordinal data. True model is the four-component mixture with $\pi = (0.14, 0.20, 0.24, 0.42)$. Downwards dendogram and unrooted dendogram in the top and bottom, respectively. Clusters of DPM locations are shown in yellow, green, red and blue

model and the proposed post-hoc clustering of its locations, recovered the true structure of the model, both in parameter values and mixture proportions. Moreover, dendograms and ordered heatmaps can be useful tool to visualize the results.

7.7 Case study: 2009-2013 Life Satisfaction in New Zealand

In this section, we estimate a finite DPM model for the Life Satisfaction in New Zealand dataset. Life satisfaction is an ordinal response with seven levels and repeated over 5 occasions on 2564 individuals ($n = 2564, p = 5, q = 7$). We fully describe the data in Chapter 2 and analyse it in Chapters 3 and 5. The main conclusions from these earlier chapters were that the correlation between ocassions needed to be taken into account and that there were at least four clusters with similar patterns over time.

Taking this into account, we use the finite DPM model proposed in this chapter (equation 7.11) to fit this dataset. The MCMC chain has 100,000 draws with a burn-in of 2000 and thinning of 50. We thus use 2000 draws for inference. Figures 7.8 and 7.9 show the results. Traceplots and marginal posteriors for all parameters could be found in Appendix 2 at the end of the chapter.

Figure 7.8, displays the posterior distribution of the non-empty groups (K). In the top panel, we can see that the most likely value for this is somewehre between 6 and 10, with a mode of 8. Given n for this dataset, this posterior distribution matches the expected value of the number of non-empty components 7.8. As expected and due to the use of the DPM prior, many of these non-empty groups are very small so when we look at the distribution of groups that have at least 13 individuals (0.5% of the sample size) in the bottom panel, we see that this changes drastically, and now we are more likely to see four and five of these relatively bigger

groups. Given that this definition of what constitutes a "big" or "small" groups is arbitrary we proceed to look at the distribution of all the locations α_i in Figure 7.9.

The top panel of Figure 7.9 shows a histogram of all estimated locations α_i as well as all the draws from the DPM prior (black dotted lines) and their expected value (red line). We can see that, the distribution for α_i is multimodal and has four well-separated components with peaks around 2.2, 4.8, 7.2 and 10 and proportions: 0.1 0.25 0.43 0.22. In other words, although the DPM selects a different number of point masses on each MCMC iteration for α_i they tend to be around the similar places forming a multimodal distribution with evident, well-separated peaks (for this dataset). We thus set $C = 4$ and post-process the DPM locations using the k-means and hierarchical clustering as outlined in Section 7.4.

The lower panel of Figure 7.9 shows a heatmap for co-clustering probabilities sorted by the above four groups using the k-means algorithm. It can be seen that the resulting classification is crisp as the co-clustering probabilities are high. Given that the clusters of DPM locations are well-separated for the NZAVS dataset this is not too surprising.

Finally, Figure 7.10 plots dendograms of the distance probability matrix obtained using hierarchical clustering. We can see that, clusters of locations are very evident (yellow, green, red and blue strips). Again, the unrooted dendongram is the most interesting as it shows that there are clearly separated clusters of locations.

Overall, the BNP model allows us to reach similar conclusions to the models presented in Chapters 3 and 5 for the NZAVS dataset: there is evidence of four latent groups, all of which have a very positive life satisfaction (all α's are positive when $\mu_1 = 0$) with stable patterns over time (occasion effects β are small in comparison to cluster effects α). Moreover, most of the population have a very positive life satisfaction as the biggest latent group, about 43% of the total, is close to the upper end of the ordinal scale.

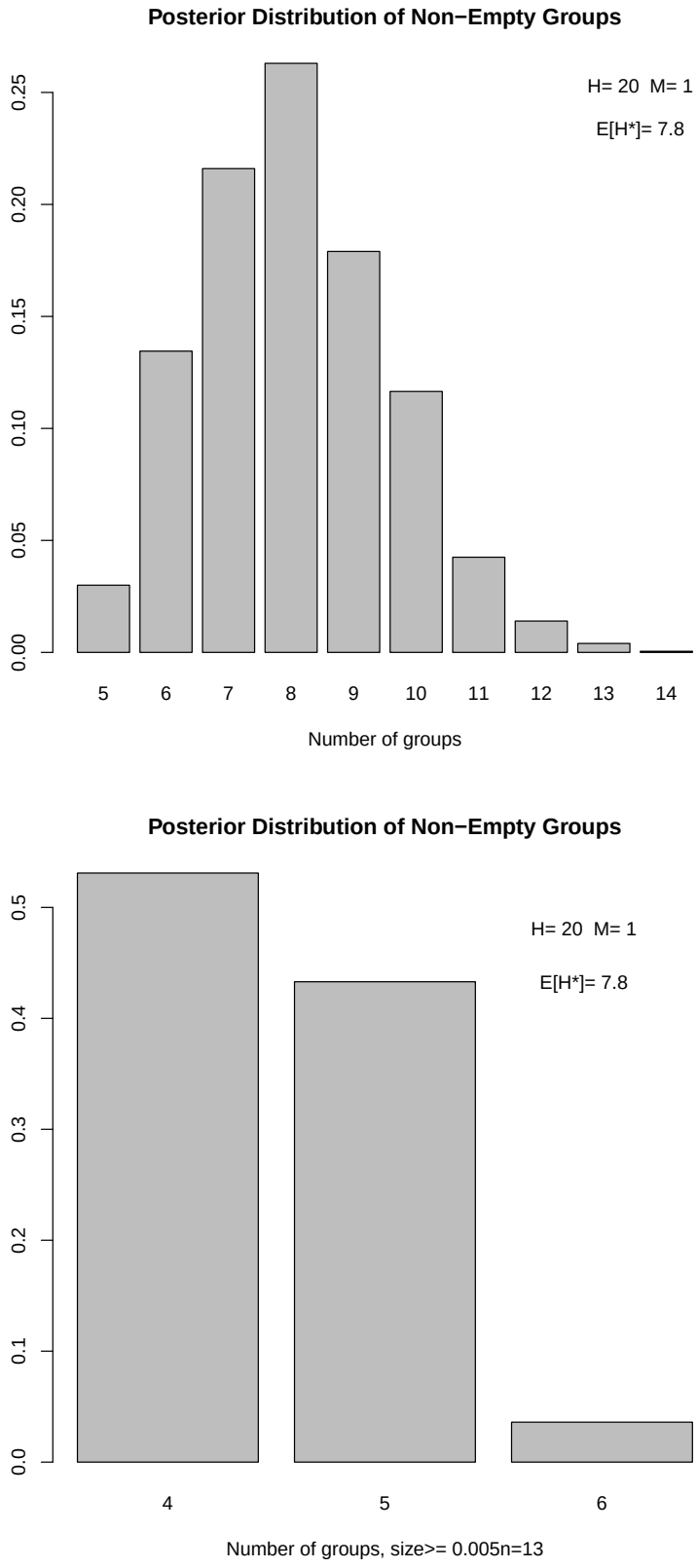


Figure 7.8: Posterior distribution for K for the DPM for the NZAVS data

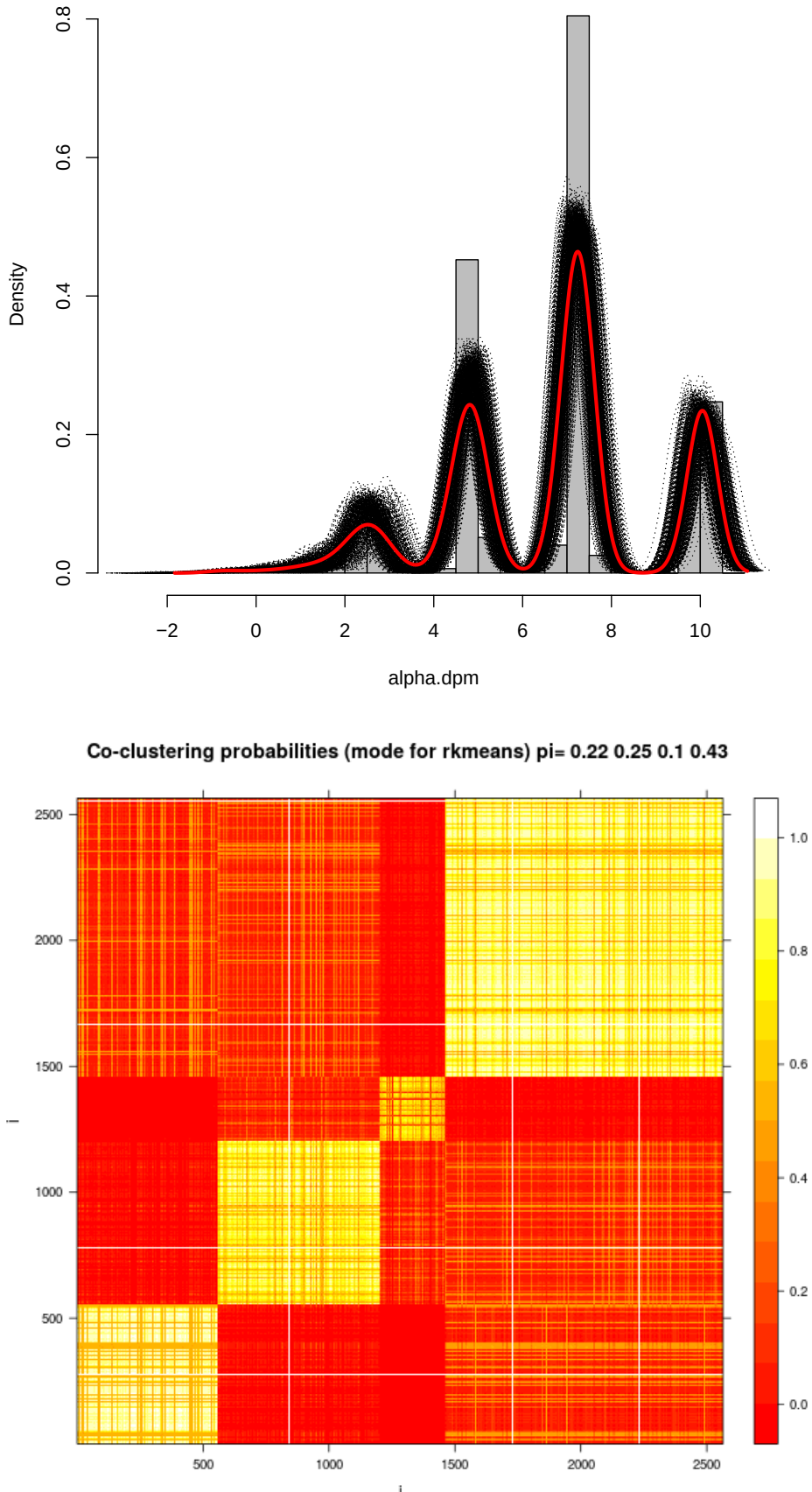


Figure 7.9: Posterior distribution of the locations (top) and ordered co-clustering probabilities (bottom) for the NZAVS dataset

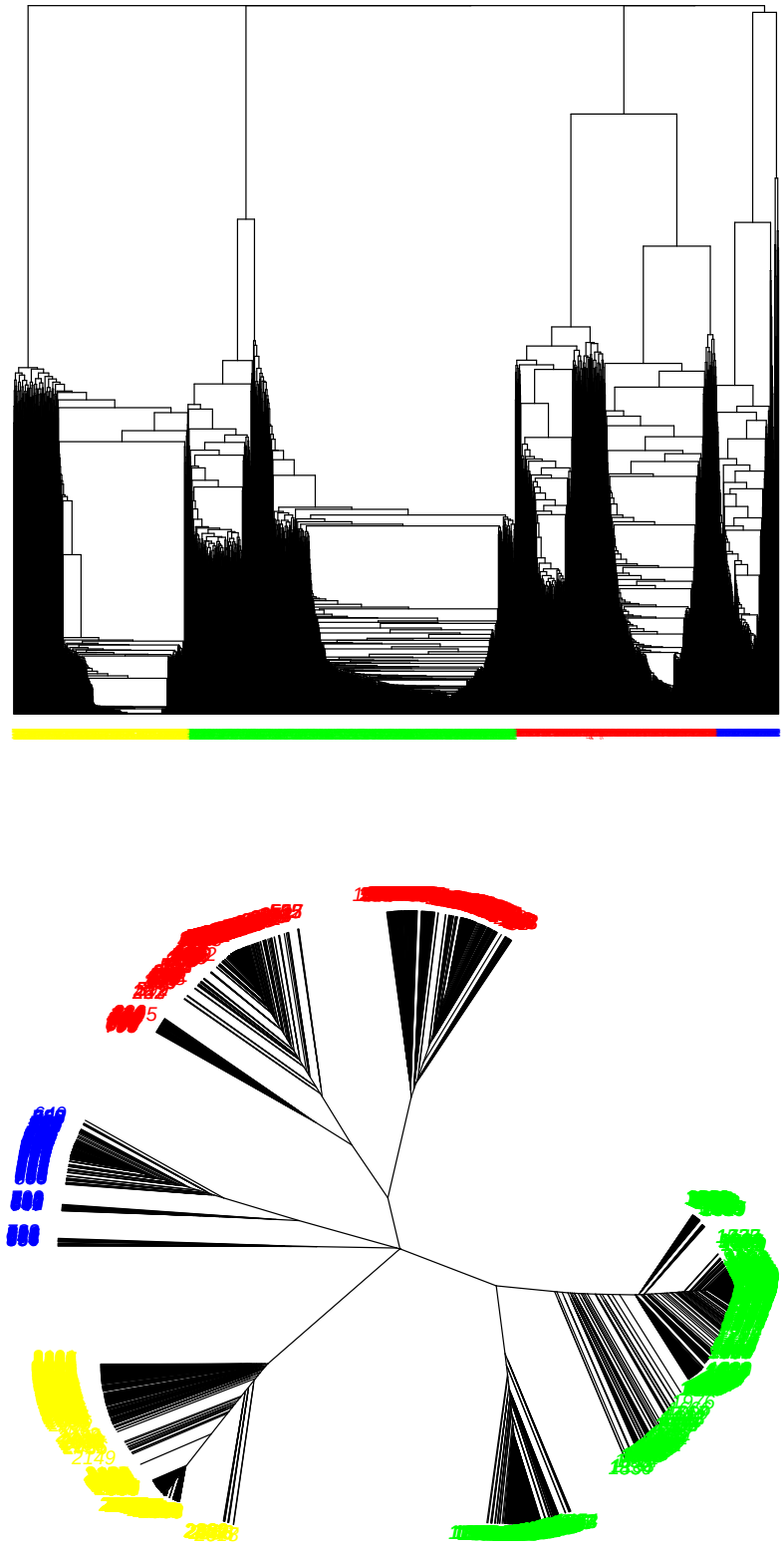


Figure 7.10: Dendograms for the NZAVS data. Downwards dendogram and unrooted dendogram in the top and bottom. Clusters of DPM locations are shown in yellow, green, red and blue

Appendix 1. Full conditionals for the DPM for repeated ordinal data

$$P(\mu, \beta, \alpha, \sigma^2_\mu, \sigma^2_\beta, \sigma^2_\alpha | Y) =$$
$$P(\mu, \beta, m, v, r, \sigma^2_\mu, \sigma^2_\beta, \sigma^2_m | Y) \propto$$
$$P(Y|\mu, \beta, m, r)$$
$$P(r|v)P(v|M)P(\mu|\sigma^2_\mu)P(\sigma^2_\mu|a_\mu, b_\mu)P(\beta|\sigma^2_\beta)P(\sigma^2_\beta|a_\beta, b_\beta)P(m|\sigma^2_m)P(\sigma^2_m|a_m, b_m)$$

Let $S_h = \{i : r_i = h\}$ and $A_h = \sum_{i=1}^n I(r_i = h)$, then:

Updates for r

$$\begin{aligned} P(r|.) \propto & P(Y|\mu, \beta, m, r)P(r|v) \\ \propto & \prod_{i=1}^{n}\prod_{j=1}^{p} P(y_{ij}|\mu, \beta_j, m_h)P(r_i|v) \end{aligned}$$

For $r_i = h$:

$$\begin{aligned} P(r_i = h|.) \propto & \prod_{j=1}^{p} P(y_{ij}|\mu, \beta_j, m_h)P(r_i = h|v) \\ \propto & \prod_{j=1}^{p} \text{Categorical}_q(y_{ij}|\mu, \beta_j, m_h)w_h \end{aligned}$$

and thus for $h = 1 \dots H$:

$$P(r_i|.) \propto \text{Categorical}_H(P(r_i = 1|.), P(r_i = 2|.), \dots, P(r_i = H|.))$$

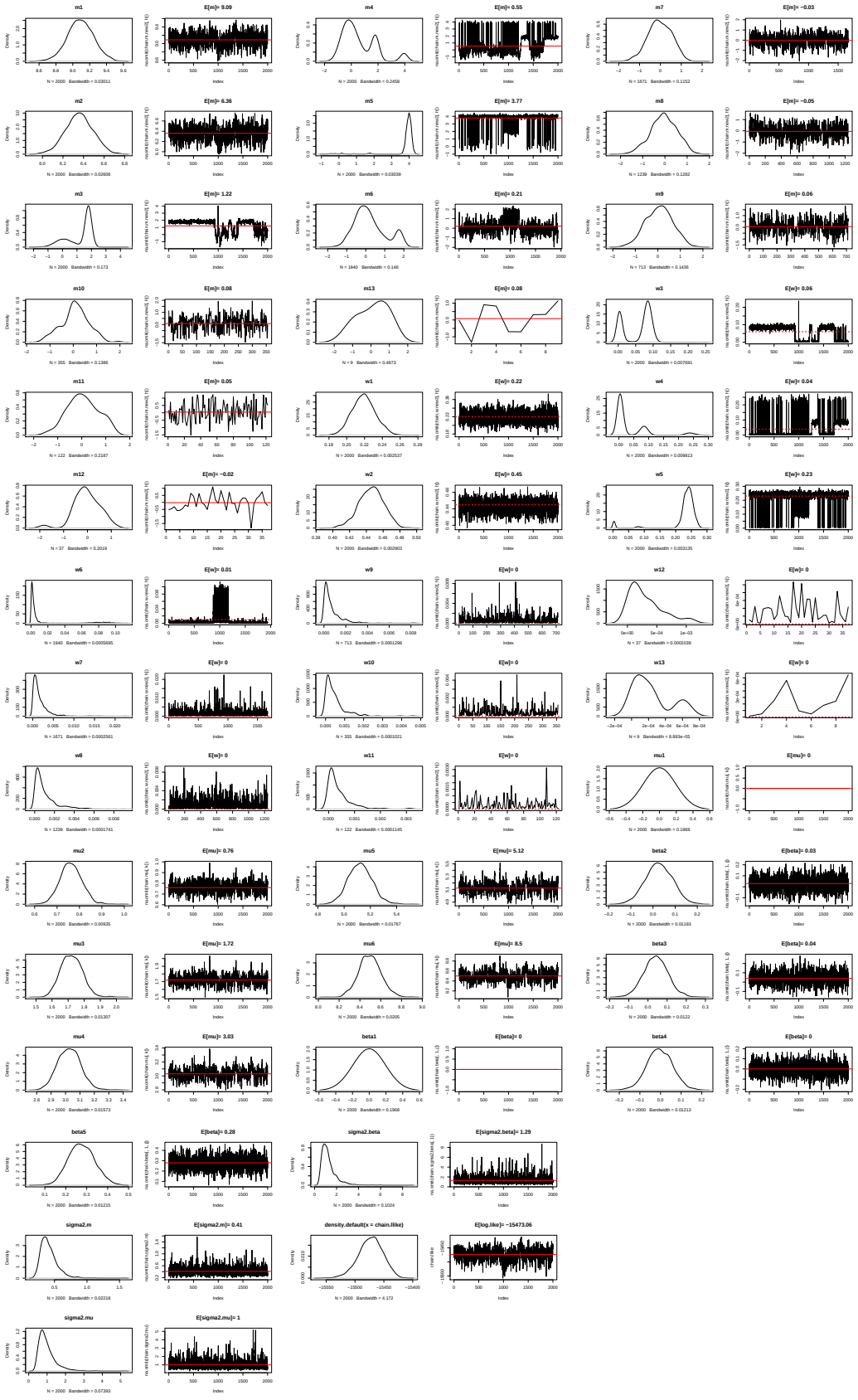
Updates for v

$$\begin{aligned} P(v|.) \propto & P(r|v)P(v|M) \\ \propto & \prod_{i=1}^{n} P(r_i|v)P(v|M) \end{aligned}$$

For each h, $v_h \overset{iid}{\sim}$ Beta$(1, M), h = 1 \dots H-1$ then:

$$\begin{aligned}
P(v_h|.) &\propto \prod_{i=1}^{n} P(r_i|v)P(v_h|M) \\
&\propto P(v_h|M) \prod_{i:r_i \geq h} P(r_i|v) \\
&\propto P(v_h|M) \prod_{i:r_i \geq h} \{v_{r_i} \prod_{l<r_i} (1-v_l)\} \\
&\propto P(v_h|M) \prod_{i:r_i = h} v_h \prod_{i:r_i > h} (1-v_h) \\
&\propto (1-v_h)^{M-1} v_h^{A_h} (1-v_h)^{B_h} \\
&\propto v_h^{A_h} (1-v_h)^{B_H+M-1} \\
&\propto \text{Beta}(A_h + 1, B_H + M)
\end{aligned}$$

Appendix 2. MCMC output for the finite DPM for repeated ordinal data estimated using the NZAVS dataset ($H = 20$)



Chapter 8

Conclusions

8.1 Summary and discussion

This PhD dissertation proposes several models to cluster repeated ordinal data using finite mixtures. In contrast to most of the existing literature, our aim is classification and not parameter estimation and thus to provide flexible and parsimonious ways to estimate latent populations and classification probabilities for repeated ordinal data.

After reviewing the relevant literature (Chapter 1) and describing the datasets (Chapter 2) used as case studies through the dissertation, we begin with simple models which assumed that conditional on the latent cluster membership observations were independent across time in Part I. Using the parametrisations of the Proportional Odds (McCullagh 1980) and Trends Odds (Capuano & Dawson 2012) models we built finite mixture models and fitted them using the EM algorithm in Chapters 3 and 4. These finite mixture models are compared using several information criteria: AIC, BIC, and ICL-BIC and applied to simulated data as well as life satisfaction and self-reported health status from the NZAVS and HILDA surveys. An important highlight of the above simulation studies was the importance of sample size to obtain meaningful estimates of the parameters when the mixture proportions are very imbalanced, that is when the

mixture model has very small clusters and big clusters.

In Part II, the repeated measurements correlation is explicitly modelled. This is done by extending the Proportional Odds model to incorporate latent random effects (Chapter 5) and latent transitional terms (Chapter 6). Following the terminology of the time-series literature, we called these approaches parameter dependent and data dependent, respectively. Models in this part are fitted using Bayesian methods to take advantage of the flexibility of MCMC methods to estimate clustering models (Frühwirth-Schnatter et al. 2012). For instance, estimation of random effects in the Bayesian paradigm, whether associated to an observed or latent variable, is an essential part of the estimation of the joint posterior distribution as both parameters and data are considered random variables. Despite being computationally more intensive than the Frequentist alternatives, Bayesian estimation is now removemore feasible thanks to the advances computer power over the last two decades.

The model proposed in Chapter 5 is a latent random effects model where the cluster effect varies over time according to a random walk with cluster-specific variance. This parametrisation provides a flexible and parsimonious ways to introduce different time patterns by cluster. Due to unavailability of the full conditional distributions in close-form, the joint posterior is simulated using a Metropolis-Hastings scheme with random-walk proposals. Model comparison is performed using the WAIC (Watanabe 2009), Bayesian information criterion for singular models such as finite mixtures. We validate the model using simulated data, and confirmed that the true values were recovered and the WAIC selected a model with the true number of mixture components. In addition to that, the WAIC allowed us to distinguished the proposed model to one with independent latent random effects (Section 5.5.2).

Chapter 6 presented a latent transitional model for repeated ordinal data. Similarly to the previous chapter, the joint posterior was simulated using a Metropolis-Hastings scheme with random-walk proposals and model

comparison carried out using the WAIC. Further to that, we also used entropy and relative entropy to compare models with different number of clusters. We validated the proposed model using simulated data and correctly identified the number of mixture components using the WAIC and the entropy measures.

Next, Chapter 7 used a Bayesian Non-Parametric approach as an alternative way to compare amongst candidate models with different number of mixture components. The use of Dirichlet Process Mixture (DPM) and the post-hoc clustering of locations provided us with a flexible way to model mixture distributions for repeated ordinal data. In a similar way to the reversible-jump MCMC (RJMCMC), Green (1995), this two-step approach allowed us to avoid the need to separately fit models with different number of mixture components and instead fit a more general encompassing model. We also presented dendograms and heatmaps and found them to be useful tools to visualize the resulting clusters.

Finally, with regard to the case studies, we saw that different information criteria selected models with different number of mixture components for all three datasets analysed: NZAVS, HILDA, and infant gut bacteria. In this respect, the predictive measures, Frequentist AIC and Bayesian WAIC, were the least parsimonious, tending to select models with higher number of parameters.The most extreme example of this was presented in Chapter 3 when clustering the NZAVS data. There, the model with the lowest AIC was a model with one parameter per each person and occasion with a total of 2613 parameters (Table 3.9). On the other hand, ICL-BIC, entropy and relative entropy; selected models with a much lower number of parameters. Guided by parsimony and interpretability, we used the models chosen by the BIC, BIC-ICL and the entropy measures in Chapters 3, 4, 5, and 6. Moreover, entropy could also be viewed as an indicator of degree of separation of the mixture components and thus provides models that are more interpretable.

8.2 Extensions and future work

There are a number ways to extend the models presented here. A first extension is the inclusion of other covariates by augmenting the linear predictor for the corresponding models. For instance, in the case of the latent transitional model, the inclusion of cluster-occasion interactions which will bring more flexibility to the model proposed in Chapter 6.Secondly, it would be interesting to incorporate missing data in the models. Approaches to handle missing data in longitudinal settings are well developed (Little & Rubin 2002) but unless the assumptions of missing completely at random or missing at random (MCAR and MAR) holds, these approaches rely on a case by case modelling of the non-ignorable drop out mechanism at hand. Recently, Skrondal & Rabe-Hesketh (2014) postulate a promising approach for transitional models for longitudinal binary data. They proposed joint working models to handle both the initial conditions problem and the missing responses. Succinctly, the unobserved response previous the first one is considered as missing an a joint working model is posed for them. Later missing responses are modelled with the same joint model with the response after the missing one being the new initial response. We plan to extent this approach to the case of repeated ordinal data.

Thirdly, extending the model to multivariate responses are also an important task for the future. In this respect, the Dependent Dirichlet Process within a Bayesian Non-Parametric approach would be an interesting avenue to explore. To date, the work in this area (Bao & Hanson 2015, DeYoreo & Kottas 2017) uses the probit link due to computational convenience. A logistic version could be developed using the approach of Holmes et al. (2006) who augmented the latent continuous representation of ordinal data with an additional layer of parameters that are iid from the Kolmogorov-Smirnov distribution.

Lastly, another important extension is to make the model scalable to

big data. Currently, models take a few hours to run on datasets with thousands of rows but estimation might become impractical with bigger datasets (millions of rows). In general, this is the case for MCMC based inference but in our case it is complicated by the unavailability of the posterior distribution in closed form and the need to simulate the joint posterior using the Metropolis-Hastings sampler. This is a technological limitation that can be alleviated by the use of grid computing and optimizing the computer code used for estimation. More importantly, however, is the development of more efficient sampling schemes that could allow faster exploration of the target distribution (or a suitable approximation). For instance, along the lines of Cowles (1996), Wainwright & Jordan (2008), Srivastava et al. (2015) joint proposals, Variational approximations and divide-and-conquer schemes could provide essential gains in this direction.

Bibliography

Agresti, A. (2013), *Categorical Data Analysis, 3rd edition*, Wiley Series in Probability and Statistics. John Wiley & Sons.

Aitkin, M. (1996), 'A general maximum likelihood analysis of overdispersion in generalized linear models', *Statistics and Computing* **6**(3), 251–262.

Aitkin, M. & Alfó, M. (1998), 'Regression models for binary longitudinal responses', *Statistics and Computing* **8**(4), 289–307.

Aitkin, M., Anderson, D. & Hinde, J. (1981), 'Statistical modelling of data on teaching styles', *Journal of the Royal Statistical Society. Series A (General)* pp. 419–461.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. Petrov & F. Csaki, eds, '2nd International Symposium on Information Theory', pp. 267–281.

Albert, J. & Chib, S. (1995), 'Bayesian residual analysis for binary response regression models', *Biometrika* **82**(4), 747–769.

Alfò, M., Salvati, N. & Ranallli, M. G. (2016), 'Finite mixtures of quantile and m-quantile regression models', *Statistics and Computing* pp. 1–24.

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

Anderson, J. & Philips, P. (1981), 'Regression, discrimination and measurement models for ordered categorical variables', *Applied Statistics* pp. 22–31.

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

Atkinson, J., Salmond, C. & Crampton, P. (2014), NZDep2013 Index of Deprivation, Technical report, Department of Public Health, University of Otago, Wellington.

Bao, J. & Hanson, T. E. (2015), 'Bayesian nonparametric multivariate ordinal regression', *Canadian Journal of Statistics* **43**(3), 337–357.

Bartholomew, D. J., Knott, M. & Moustaki, I. (2011), *Latent Variable Models and Factor Analysis: A Unified Approach*, John Wiley & Sons.

Bartolucci, F. (2006), 'Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **68**(2), 155–178.

Bartolucci, F., Bacci, S. & Pennoni, F. (2014), 'Longitudinal analysis of self-reported health status by mixture latent auto-regressive models', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **63**(2), 267–288.

Bartolucci, F. & Farcomeni, A. (2009), 'A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure', *Journal of the American Statistical Association* **104**(486), 816–831.

Bartolucci, F., Farcomeni, A. & Pennoni, F. (2012), *Latent Markov Models for Longitudinal Data*, CRC Press.

Bates, G. E. & Neyman, J. (1952), Contributions to the theory of accident proneness II. True or false contagion, Technical report, University of California publications in Statistics.

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

Blackwell, D. & MacQueen, J. B. (1973), 'Ferguson distributions via Pólya urn schemes', *The Annals of Statistics* pp. 353–355.

Bock, R. D. & Jones, J. (1968), *The Measurement and Prediction of Judgment and Choice*, Oxford. England.

Cappé, O., Moulines, E. & Rydén, T. (2005), *Inference in Hidden Markov Models*, Springer.

Capuano, A. & Dawson, J. (2012), 'The Trend Odds model for ordinal data', *Statistics in Medicine* **32**, 2250–2261.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M. et al. (2006), 'Deviance information criteria for missing data models', *Bayesian analysis* **1**(4), 651–673.

Chan, E. et al. (2015), Unpublished data. Presentation given by Aaron Darling. 2015 BioInfoSummer, University of Sydney, Australia.

Cheon, K., Thoma, M. E., Kong, X. & Albert, P. S. (2014), 'A mixture of transition models for heterogeneous longitudinal ordinal data: with applications to longitudinal bacterial vaginosis data', *Statistics in Medicine* **33**(18), 3204–3213.

Chib, S. & Greenberg, E. (1995), 'Understanding the Metropolis-Hastings algorithm', *The American Statistician* **49**(4), 327–335.

Cowles, M. K. (1996), 'Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models', *Statistics and Computing* **6**(2), 101–111.

Cox, D. R. (1972), 'Models and life-tables regression', *Journal of the Royal Statistical Society* **34**, 187–220.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton university press.

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

Dahl, D. B. (2006), Model-based clustering for expression data via a Dirichlet process mixture model, *in* 'Bayesian Inference for Gene Expression and Proteomics.', Cambridge University Press, pp. 201–218.

De Iorio, M., Müller, P., Rosner, G. L. & MacEachern, S. N. (2004), 'An ANOVA model for dependent random measures', *Journal of the American Statistical Association* **99**(465), 205–215.

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

DeYoreo, M. & Kottas, A. (2017), 'Bayesian nonparametric modeling for multivariate ordinal regression', *Journal of Computational and Graphical Statistics* (to-appear).

Diggle, P. J., Heagerty, P. J., Liang, K.-Y. & Zeger, S. L. (2002), *Analysis of Longitudinal Data, 2nd Edition*, Oxford University Press.

Everitt, B. & Hand, D. (1981), Mixtures of discrete distributions, *in* 'Finite Mixture Distributions', Springer, pp. 89–105.

Fahrmeir, L. & Tutz, G. (2001), *Multivariate Statistical Modelling based on Generalized Linear Models, 2nd edition*, Springer New York.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problems', *The Annals of Statistics* pp. 209–230.

Ferguson, T. S. (1983), 'Bayesian density estimation by mixtures of normal distributions', *Recent advances in Statistics* **24**(1983), 287–302.

Fernández, D., Arnold, R. & Pledger, S. (2016), 'Mixture-based clustering for the ordered stereotype model', *Computational Statistics and Data Analysis* .

Fernández, D. & Pledger, S. (2015), 'Categorising count data into ordinal responses with application to ecological communities', *Journal of Agricultural, Biological, and Environmental Statistics* pp. 1–15.

Fernández, D., S., P., Arnold, R., Liu, I. & Costilla, R. (2017), 'Mixture-based clustering for binomial, count and ordinal data:a summary of recent developments', *Advances in Data Analysis and Classification* .

Fonseca, J. R. & Cardoso, M. G. (2007), 'Mixture-model cluster analysis using information theoretical criteria', *Intelligent Data Analysis* **11**(2), 155–173.

Fraley, C. & Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American Statistical Association* **97**(458), 611–631.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer Science and Business Media.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A. & Winter-Ebmer, R. (2012), 'Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering', *Journal of Applied Econometrics* **27**(7), 1116–1137.

Frydman, H. (2005), 'Estimation in the mixture of Markov chains moving with different speeds', *Journal of the American Statistical Association* **100**(471), 1046–1053.

Geisser, S. & Eddy, W. F. (1979), 'A predictive approach to model selection', *Journal of the American Statistical Association* **74**(365), 153–160.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014), *Bayesian Data Analysis, 3rd edition*, Taylor & Francis.

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(4), 457–472.

Ghosh, J. K. & Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, Springer.

Goodman, L. A. (1979), 'Simple models for the analysis of Association in cross-classifications having ordered categories', *Journal of the American Statistical Association* **74**(367), 537–552.

Goodman, L. A. & Kruskal, W. H. (1954), 'Measures of Association for cross classifications', *Journal of the American Statistical Association* **49**, 732–764.

Govaert, G. & Nadif, M. (2005), 'An EM algorithm for the block mixture model', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**(4), 643–647.

Govaert, G. & Nadif, M. (2008), 'Block clustering with bernoulli mixture models: comparison of different approaches', *Computational Statistics and Data Analysis* **52**, 3233–3245.

Govaert, G. & Nadif, M. (2010), 'Latent block model for contingency table', *Communications in Statistics. Theory and Methods* **39**(3), 416–425.

Green, P. J. (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82**(4), 711–732.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning, 2nd edition*, Springer.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.

Heckman, J. J. (1981*a*), Heterogeneity and state dependence, *in* S. Rosen, ed., 'Studies in Labor Markets', University of Chicago Press.

Heckman, J. J. (1981*b*), The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, *in* C. F. Manski, D. McFadden et al., eds, 'Structural Analysis of Discrete Data with Econometric Applications', University of Chicago Press.

Heiss, F. (2008), 'Sequential numerical integration in nonlinear state space models for microeconometric panel data', *Journal of Applied Econometrics* **23**(3), 373–389.

Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G. (2010), *Bayesian Nonparametrics*, Vol. 28, Cambridge University Press.

Holmes, C. C., Held, L. et al. (2006), 'Bayesian auxiliary variable models for binary and multinomial regression', *Bayesian Analysis* **1**(1), 145–168.

Ishwaran, H. & James, L. F. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. & James, L. F. (2002), 'Approximate Dirichlet process computing in finite normal mixtures', *Journal of Computational and Graphical Statistics* pp. 508–532.

Ishwaran, H. & Zarepour, M. (2000), 'Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models', *Biometrika* **87**, 371–390.

Jara, A., Quintana, F. & San Martín, E. (2008), 'Linear mixed models with skew-elliptical distributions: A Bayesian approach', *Computational Statistics and Data Analysis* **52**(11), 5033–5045.

Johnson, V. E. & Albert, J. (1999), *Ordinal Data Modeling*, Statistics for Social Science and Public Policy. Springer-Verlag New York.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Kaufman, L. & Rousseeuw, P. J. (1990), *Finding groups in data: An introduction to cluster analysis*, Wiley New York.

Kedem, B. & Fokianos, K. (2005), *Regression models for time series analysis*, Vol. 488, John Wiley & Sons.

Kendall, M. G. (1945), 'The treatment of ties in rank problems', *Biometrika* **33**, 239–251.

Korwar, R. M. & Hollander, M. (1973), 'Contributions to the theory of Dirichlet processes', *The Annals of Probability* pp. 705–711.

Labiod, L. & Nadif, M. (2011), Co-clustering for binary and categorical data with maximum modularity, *in* '2011 IEEE 11th International Conference on Data Mining (ICDM)', IEEE, pp. 1140–1145.

Lazarsfeld, P. (1950), The logical and mathematical foundations of latent structure analysis, *in* S. Stouffer, L. Guttman, E. Suchman, L. P., S. Star & J. Clausen, eds, 'Studies in Social Psychology in World War II. Vol IV. Measurement and Prediction', Princeton University Press.

Lewis, S. J. G., Foltynie, T., Blackwell, A. D., Robbins, T. W., Owen, A. M. & Barker, R. A. (2003), 'Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach', *Journal of Neurology, Neurosurgery and Psychiatry* **76**, 343–348.

Lipsitz, S. R., Kim, K. & Zhao, L. (1994), 'Analysis of repeated categorical data using generalized estimating equations', *Statistics in Medicine* **13**(11), 1149–1163.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data, 2nd edition*, J. Wiley.

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

MacDonald, I. L. & Zucchini, W. (1997), *Hidden Markov and Other Models for Discrete-Valued Time Series*, Vol. 110, CRC Press.

MacEachern, S. N. (1999), Dependent nonparametric processes, *in* 'ASA proceedings of the section on Bayesian Statistical science', Alexandria, Virginia. Virginia: American Statistical Association; 1999, pp. 50–55.

Manly, B. F. (2005), *Multivariate Statistical Methods: A Primer*, CRC Press.

Marin, J.-M., Mengersen, K. & Robert, C. P. (2005), 'Bayesian modelling and inference on mixtures of distributions', *Handbook of Statistics* **25**(16), 459–507.

Matechou, E., Liu, I., Fernández, D., Farias, M. & Gjelsvik, B. (2016), 'Biclustering models for two-mode ordinal data', *Psycometrika* pp. 611–624.

McCullagh, P. (1980), 'Regression models for ordinal data', *Statistical Methodology* **42**, 109–142.

McCulloch, C., Searle, S. & Neuhaus, J. (2008), *Generalized, Linear, and Mixed Models*, John Wiley & Sons.

McLachlan, G. & Peel, D. (2000), *Finite Mixture Models*, Wiley Series in Probability and Statistics.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The Journal of Chemical Physics* **21**(6), 1087–1092.

Mitra, R. & Müller, P. (2015), *Nonparametric Bayesian Methods in Biostatistics and Bioinformatics*, Springer-Verlag.

Molitor, J., Papathomas, M., Jerrett, M. & Richardson, S. (2010), 'Bayesian profile regression with an application to the national survey of children's health', *Biostatistics* pp. 484–498.

Müller, P., Quintana, F., Jara, A. & Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, Springer Series in Statistics. Springer International Publishing.

Naylor, J. C. & Smith, A. F. (1982), 'Applications of a method for the efficient computation of posterior distributions', *Applied Statistics* pp. 214–225.

Nylund, K. L., Asparouhov, T. & Muthén, B. O. (2007), 'Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study', *Structural Equation Modeling* **14**(4), 535–569.

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

Peterson, B. & Harrell, F. (1990), 'Partial proportional odds models for ordinal response variables', *Applied Statistics* **39**, 205–217.

Phadia, E. G. (2013), *Prior Processes and Their Applications: Nonparametric Bayesian Estimation*, Springer Science & Business Media.

Pinheiro, J. C. & Bates, D. M. (1995), 'Approximations to the log-likelihood function in the nonlinear mixed-effects model', *Journal of Computational and Graphical Statistics* **4**(1), 12–35.

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

Pledger, S. & Arnold, R. (2014), 'Clustering, scaling and correspondence analysis: unified pattern-detection models using mixtures', *Computational Statistics and Data Analysis* **71**, 241–261.

Richardson, S. & Green, P. J. (1997), 'On Bayesian analysis of mixtures with an unknown number of components', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* pp. 731–792.

Robert, C. P. & Casella, G. (2005), *Monte Carlo Statistical Methods*, Springer Texts in Statistics. Springer-Verlag New York.

Roberts, G. O., Gelman, A., Gilks, W. R. et al. (1997), 'Weak convergence and optimal scaling of random walk metropolis algorithms', *The Annals of Applied Probability* **7**(1), 110–120.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Sethuraman, J. (1994), 'A constructive definition of Dirichlet priors', *Statistica Sinica* pp. 639–650.

Si, Y., Liu, P., Li, P. & Brutnell, T. P. (2014), 'Model-based clustering for RNA-Seq data', *Bioinformatics* pp. 197–205.

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, CRC Press.

Skrondal, A. & Rabe-Hesketh, S. (2014), 'Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **63**(2), 211–237.

Snell, E. (1964), 'A scaling procedure for ordered categorical data', *Biometrics* **20**(3), 592–607.

Somers, R. H. (1962), 'A new assymetric measure of Association for ordinal variables', *American Sociological Review* **27**, 799–811.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Linde, A. (2014), 'The deviance information criterion: 12 years on', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76**(3), 485–493.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**(4), 583–639.

Srivastava, S., Li, C. & Dunson, D. B. (2015), 'Scalable bayes via barycenter in wasserstein space', *arXiv preprint arXiv:1508.05880* .

Stephens, M. (2000), 'Dealing with label switching in mixture models', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **62**, 795–809.

Thompson, R. & Baker, R. (1981), 'Composite link functions in generalized linear models', *Applied Statistics* pp. 125–131.

Toledano, A. Y. & Gatsonis, C. (1996), 'Ordinal regression methodology for roc curves derived from correlated data', *Statistics in Medicine* **15**(16), 1807–1826.

Tutz, G. & Hennevogl, W. (1996), 'Random effects in ordinal regression models', *Computational Statistics and Data Analysis* **22**(5), 537–557.

Vermunt, J. K. & Hagenaars, J. A. (2004), Ordinal longitudinal data analysis, *in* R. Hauspie, N. Cameron & L. Molinari, eds, 'Methods in Human Growth Research', Cambridge University Press.

Vermunt, J. K., Langeheine, R. & Bockenholt, U. (1999), 'Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates', *Journal of Educational and Behavioral Statistics* **24**(2), 179–207.

Vermunt, J. K. & Magidson, J. (2000), 'Latent Gold Users Guide', *Belmont, MA: Statistical Innovations Inc* .

Vermunt, J. K. & Van Dijk, L. (2001), 'A nonparametric random-coefficients approach: The latent class regression model', *Multilevel Modelling Newsletter* **13**(2), 6–13.

Wainwright, M. & Jordan, M. (2008), *Graphical Models, Exponential Families, and Variational Inference*, Foundations and Trends in Machine Learning, Now Publishers.

Walker, S. H. & Duncan, D. B. (1967), 'Estimation of the probability of an event as a function of several independent variables', *Biometrika* **54**(1-2), 167–179.

Watanabe, S. (2009), *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press.

Watanabe, S. (2010), 'Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory', *The Journal of Machine Learning Research* **11**, 3571–3594.

Wedel, M. & DeSarbo, W. S. (1994), A review of recent developments in latent class regression models, *in* R. P. Bagozzi, ed., 'Advanced Methods of Marketing Research', Blackwell Cambridge, MA.

Wedel, M. & DeSarbo, W. S. (1995), 'A mixture likelihood approach for generalized linear models', *Journal of Classification* **12**(1), 21–55.

Wiggins, L. M. (1973), *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*, Jossey-Bass/Elsevier international series, Elsevier Scientific Publishing Company.

Wilkinson, L. & Friendly, M. (2009), 'The history of the cluster heat map', *The American Statistician* **63**(2).

Wooldridge, J. M. (2005), 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity', *Journal of Applied Econometrics* **20**(1), 39–54.

Zucchini, W. & MacDonald, I. L. (2009), *Hidden Markov Models for Time Series: An Introduction using R*, CRC press.