VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wānanga o te Ūpoko o te Ika a Māui

School of Mathematics, Statistics
and Operations Research
Te Kura Mātai Tatauranga, Rangahau Pūnaha

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341
Fax: +64 4 463 5045
Email: office@msor.vuw.ac.nz

Cluster analysis of repeated ordinal data: A model approach based on finite mixtures

Roy Costilla

Supervisors: Ivy Liu and Richard Arnold

May 16, 2014

PhD Proposal

1 Introduction

A variable with an ordered categorical scale is called *ordinal* (Agresti, 2010). That is, ordinal data is categorical data where outcome categories have a logical order and thus the order of the categories 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 a Likert-like scale. Conversely, a categorical variable with an unordered scale is called *nominal*. In this case, categories differ in quality not in quantity (Agresti, 2010). Religious affiliation (Non-religious, Christian, Muslim, Jewish, Buddhist, Other), geographical location (North, East, South, West), preferred method of commuting (bus, train, bike, walk, other) and are examples of nominal variables.

Analyses of ordinal data are very common but often don't 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 cluster approaches such as 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't 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 overtime (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.

This proposal develops cluster models based on finite mixtures to attempt to fill some of these gaps. We use cumulative logits, probability models based on likelihoods, and thus provide fuzzy clustering models in which observations could come from any latent cluster with some probability. In particular, we have developed two models for two-way data, the Trend Odss model (TOM) and the Proportional Odds Model (POM), and will extent it to three-way data and joint models.

The structure of this document is as follows. Section 2 reviews the existing literature for ordinal data, repeated measures in ordinal data, and model based cluster analysis. Next, section 3 shows the methodology in detail for the two models developed, including their likelihoods, presenting estimation strategies, model selection, and visualisation tools. Section 4 presents work we have carried out so far, using longitudinal data from an existing

survey (the Household, Income and Labour Dynamics in Australia -HILDA) as well as simulated data. This work has been being presented two local conferences and will be presented in an international conference in July. Finally, sections 5 and 6 detail the research goals and the plan to carry out the research.

2 Literature Review

2.1 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 develop after the influential articles by McCullagh1980 on modelling of cumulative probabilities using a logit link and Goodman (1979) on loglinear models for odd ratios of ordered categories. Substantial developments have been made since then and are well documented elsewhere (Liu & Agresti, 2005; Agresti, 2010, 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 (POM) by McCullagh (1980) is a cumulative logit model and is the most popular model to analyse 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 $\mu_1 < \mu_2 \cdots < \mu_q$. These parameters are the cut-off 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.β captures the effect of the predictors on the cumulative probabilities and is the same for all the cumulative probabilities (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 preceding β, 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 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 ($k = 1 \ldots 5$) and differs only in its location.

Alternatively, the POM has also a latent variable representation (J. 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 $\epsilon \sim \text{Logistic}(0, \pi^2/3)$. Figure 2 shows a graphical representation, the ordinal response Y (left 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 and is the same for all the ordinal categories.

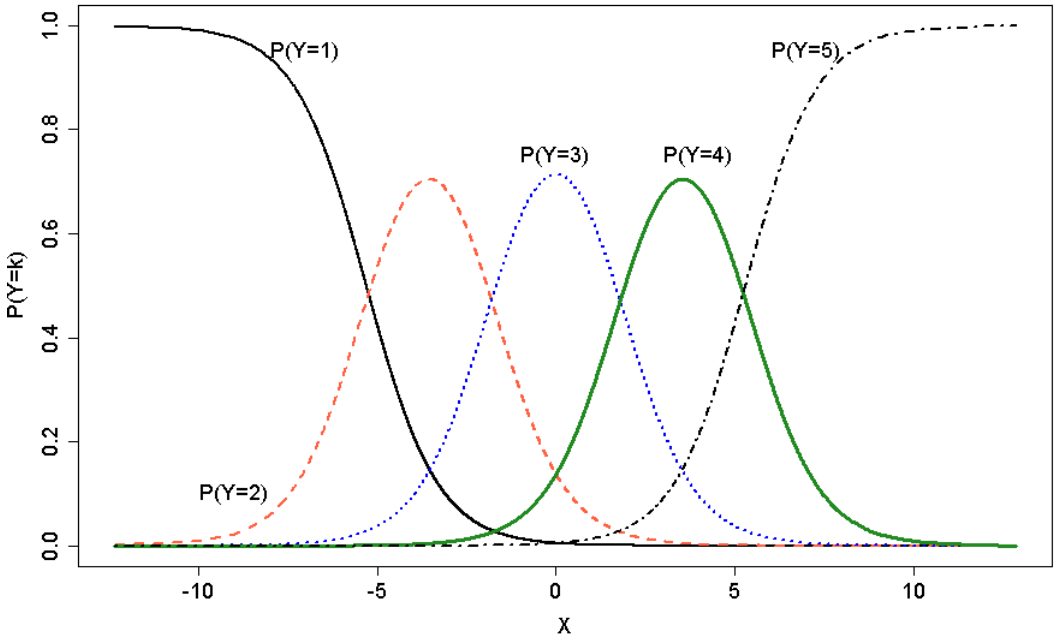


Figure 1: Individual category probabilities for the POM with five response categories

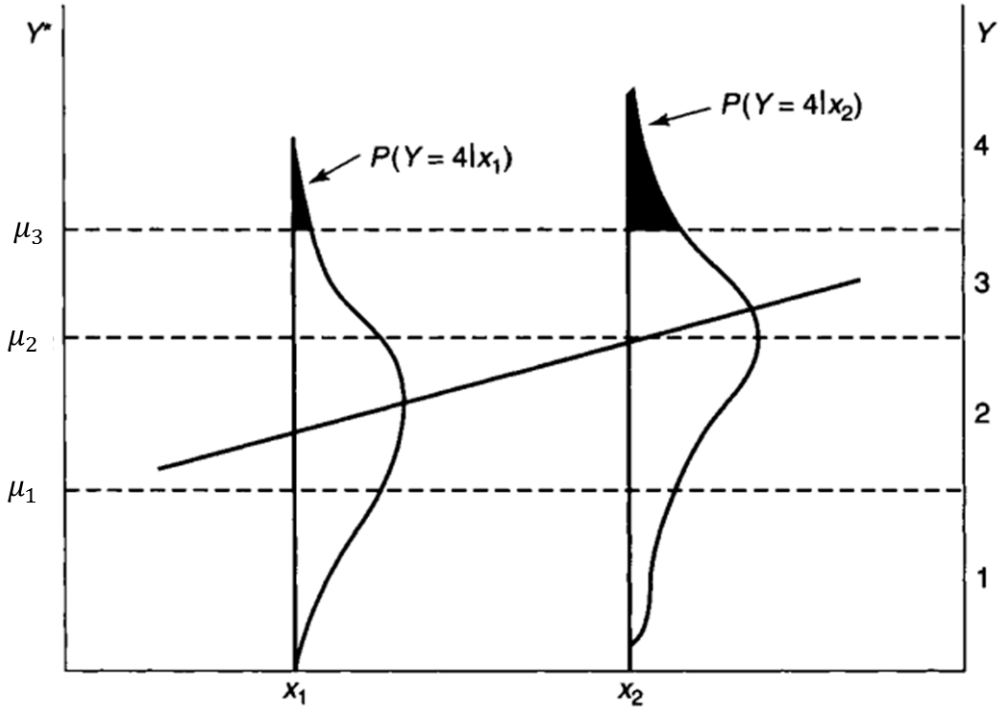


Figure 2: Latent variable representation for POM, reprinted from Agresti (2010)

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 describe above. Walker & Duncan (1967); McCullagh (1980) proposed the Fisher scoring al-

gorithm, an iteratively reweighted least squares algorithm, for this task. A sufficiently large n guarantees a global maximum but a finite n doesn't. In the latter, the ML function may exhibit local maxima or not have one at all (McCullagh, 1980). Matechou et al. (2014) extented the POM to perform model based cluster analysis, section 2.3.

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 adding 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 to make sure the cumulative probabilities are non-decreasing. Intrinsically, therefore the TOM is a constraint 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 3 shows a graphical representation of the TOM for five response categories and one continuous predictor. In contrast to the POM, figure 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 not longer mirror images of each other.

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 J. A. 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.Fernandez et al. (2014) extend the Stereotype model to perform model based cluster analysis, see section 2.3. 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).

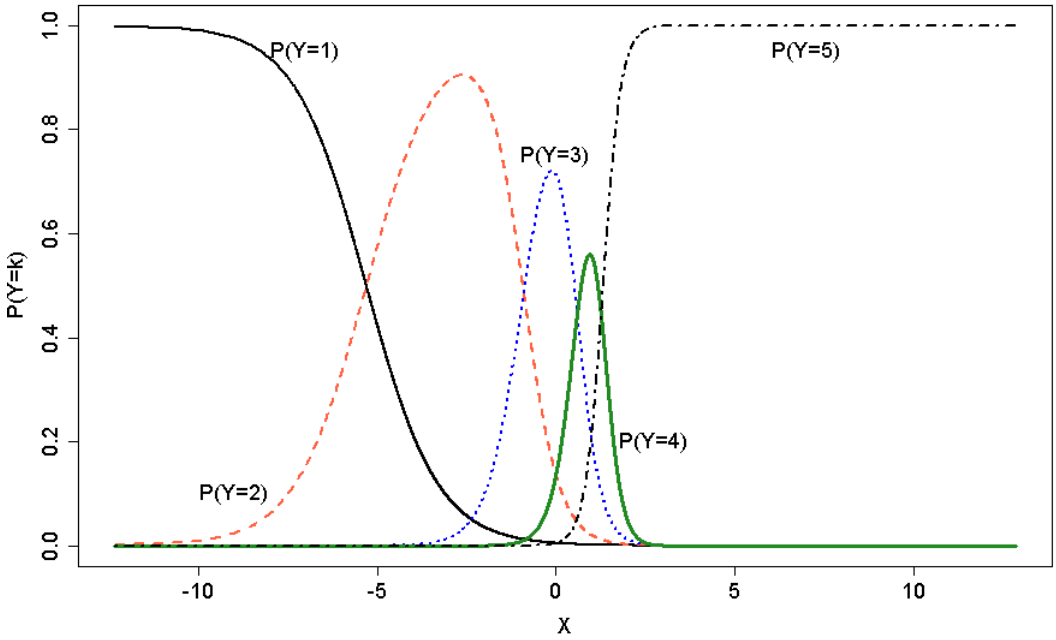


Figure 3: Individual category probabilities for the TOM with five response categories

2.2 Models for Repeated Ordinal Data

Repeated ordinal data arise when an ordinal response is recorded at various occasions for each subject, such as in longitudinal studies. We next discuss three main approaches to analyse such data: marginal models, subject-specific models and transitional models. Agresti (2010); Vermunt & Hagenaars (2004) provide a full treatment of the topic.

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_t in $(Y_1, Y_2, \ldots, Y_T)$ be an ordinal response with q categories. The marginal model with cumulative logit link has the form

$$\text{Logit}[P(Y_t \leq k|x_t)] = \mu_k - \beta' x_t \qquad k = 1, \ldots, q-1; t = 1, \ldots, T$$

Where $x_t = (x_{1t}, x_{2t}, \ldots, x_{mt})$ contains the values of the m predictors at observation t. Model fitting is mostly performed using a generalised estimating equations (GEE) approach, and not ML. The GEE method is a multivariate generalisation of quasi likelihood that only specifies the marginal regression models, as in the equation above, and a (guess for a) working correlation structure among the T responses. Lipsitz et al. (1994) and Toledano & Gatsonis (1996) presented cumulative logit and probit models for repeated ordinal responses.

Importantly, marginal models focus on the dependence of their T (first-order) marginal distribution on the explanatory variables and leave aside the multivariate dependence among the T repeated responses. They treat the joint dependence structure as nuisance. Given that our aim is to model explicitly the latter, we won't be using this type of model.

In contrast to that, subject-specific models are conditional models that describe effects at the individual or unit level. They are also known as random-effects, cluster-specific or multi-level models and jointly model the distributions of the response and the individual effects. Random effects models belong to the class of multivariate generalised linear mixed models (Multivariate GLMM) when the response has a distribution in the exponential fam-

ily. The individual effects are assumed to follow a certain 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

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

where $a_i \sim N(0, \sigma^2)$. 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 models 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 (McCulloch et al., 2008; Tutz & Hennevogl, 1996) since that the more terms for the random effects the more multiple integrals needed to be solve numerically. Higher-dimensional integrals are solve through Monte Carlo simulation or pseudo-likelihood methods (Liu & Agresti, 2005). Bayesian approaches are however natural in these cases, and provide an attractive way forward. Section 3.2 develops a cumulative logit with random effects by latent cluster and occasion estimated 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 (2002) used a cumulative logit transitional model in the context of a longitudinal medical study.

As remarked by Liu & Agresti (2005), the use of any of these three approaches depends on the problem at hand, that is whether interpretation 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 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 ordinal response depends on how many previous observations are include in the model. Also the effects of the other explanatory variables diminish markedly (Agresti, 2010). In addition to that, effects in a cluster-specific model are larger in magnitude than those in a population-averaged model.

2.3 Model-based cluster analysis for ordinal data

Traditional cluster analysis approaches treat ordinal responses as continuos and reduce the dimensionalty 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't 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 ad hoc distance metrics and crude similarity measures and thus don't 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.

In addition to that, model-based clustering methods using finite mixtures have been proposed by several authors (McLachlan & Peel, 2000; Everitt et al., 2001). See a recent literature review by Melnykov & Maitra (2010). This approach poses probabilistic models using finite mixtures and carries out fuzzy clustering for the rows or columns but not for both simultaneously e.g. one-way clustering. In our case, individuals are represented in the rows and occasions of the ordinal response in the columns. Models are fitted using the Expected-Maximisation algorithm (EM) (Dempster et al., 1977) and focus on either continuous, discrete or nominal responses. A major advantage of this approach it the availability of likelihoods, for the probability models, and therefore access to various model selection criteria to evaluate and compare different models. Although named differently Latent Class Models (LC) is a closely related model-based approach that could be used to cluster ordinal responses, as well as continuous and categorical data. LC also perfoms one-way clustering, mostly of the rows given their aim to find latent groups among individuals or subjects, using finite mixtures. In contrast to random-effects models reviewed before, section 2.2 , LC is a non-parametric random-effects approach since the distribution of the random effects is assumed to be multinomial, e.g. random effects don't come from any parametric family. Models also are fitted using the EM algorithm and compared using a likelihood-based goodness of fit criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Of note in this approach are Skrondal & Rabe-Hesketh (2004) and Magidson & Vermunt (2004) implemented this approach and made them available in standard software, GLLAMM (Generalised Linear Latent Mixed Models) in the statistical packages Stata and Latent Gold. An important caveat of LC models is that exclusive reliance on AIC and BIC for model comparison might not be appropiate for finite mixtures as it tends to overestimate the number of latent clusters. Section 3.1.5 provides more details and section 5 lists our research goals in this area.

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. (2014) and Fernandez et al. (2014) have extended these models to ordinal responses. The former used the proportional odds and the latter the Stereotype model and enable them to handle row, column and biclustered data. In turn, the purpose of this research is to extent these models to the case of repeated ordinal data as details in coming sections 3 and 5.

3 Model-based cluster analysis for repeated ordinal data

This section presents some initial work we have carried out on the project so far. It is been presented at two local conferences: the 2013 New Zealand Statistical Association's annual meeting and the 2014 Research Symposium of the School of Economics and Finance of our university.

3.1 Trend Odds Model with clustering

In this section, we extend the TOM to the case of a finite number of clusters. The setup that follows will be used throughout the proposal. Let data $\mathbf{Y}$ 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 $k = 1, ..., q$. We finally define an indicator variable $I(y_{ij} = k)$ equal to 1 if the condition $y_{ij} = k$ is satisfied and 0 otherwise.

3.1.1 Row-clustering

We start with the case of row-clustering. Rows are assumed to come from any of the $r = 1, \ldots, R$ row groups with proportions $\pi_1, \ldots, \pi_R$. 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 $P(y_{ij} = k|i \in r) = \theta_{rjk}$. The linear predictor case is then:

$$\text{Logit}[P(y_{ij} \leq k|i \in r)] = \mu_k - \alpha_r - \gamma_r t_k \tag{1}$$

or equivalently

$$\theta_{rjk} = \frac{\exp(\mu_k - \alpha_r - \gamma_r t_k)}{1 + \exp(\mu_k - \alpha_r - \gamma_r t_k)} - \frac{\exp(\mu_{k-1} - \alpha_r - \gamma_r t_{k-1})}{1 + \exp(\mu_{k-1} - \alpha_r - \gamma_r t_{k-1})} \tag{2}$$

Where $\alpha_1 = \gamma_1 = 0$. μ_k is the k^{th} cut-off point, α_r is the effect of row-cluster r and $\gamma_r t_k$ represents any non-proportional odds for the row-cluster since it depends both on r and k. Setting $t_k = k - 1$, following Capuano & Dawson (2012), the constraint

$$\mu_k - \mu_{k-1} \geq \gamma_r \qquad \forall r, k \tag{3}$$

is also necessary to make sure the cumulative probabilities are non-decreasing in the ordinal outcomes.

As an example, Figure 4 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 $\gamma = (0, -2, -1)$. As expected, in the case of the POM θ_{rjk} have the same shape for all clusters but different location (α_r). In contrast, for the TOM θ_{rjk} have both different shape (γ) and location (α_r) in all clusters.

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

$$L(\phi, \pi|\mathbf{Y}) = \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)} \tag{4}$$

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 (summarised in π_r) are unknown. The number of model parameters is equal to: $v = (q - 1) + 3(R - 1)$

3.1.2 Column-clustering

Column-clustering is the same as the row-clustering, a one-way clustering, with the transposed data. That is, we could obtained column groups by exchanging row and columns and applying the above row-cluster model. Setting C as the number of mixture components, κ_c as the mixture proportion for group c, $P(y_{ij} = k|j \in c) = \theta_{ick}$, and β_c and δ_c as the cluster effects, the linear predictor is

$$\text{Logit}[P(y_{ij} \leq k|j \in c)] = \mu_k - \beta_c - \delta_c t_k \tag{5}$$

Note also that $C < p$ and $\sum_{c=1}^{C} \kappa_c = 1$, $\beta_1 = \delta_1 = 0$. Making $t_k = k - 1$ as before, the non-decreasing cumulative probabilities constraint is $\mu_k - \mu_{k-1} \geq \delta_c \forall c, k$.

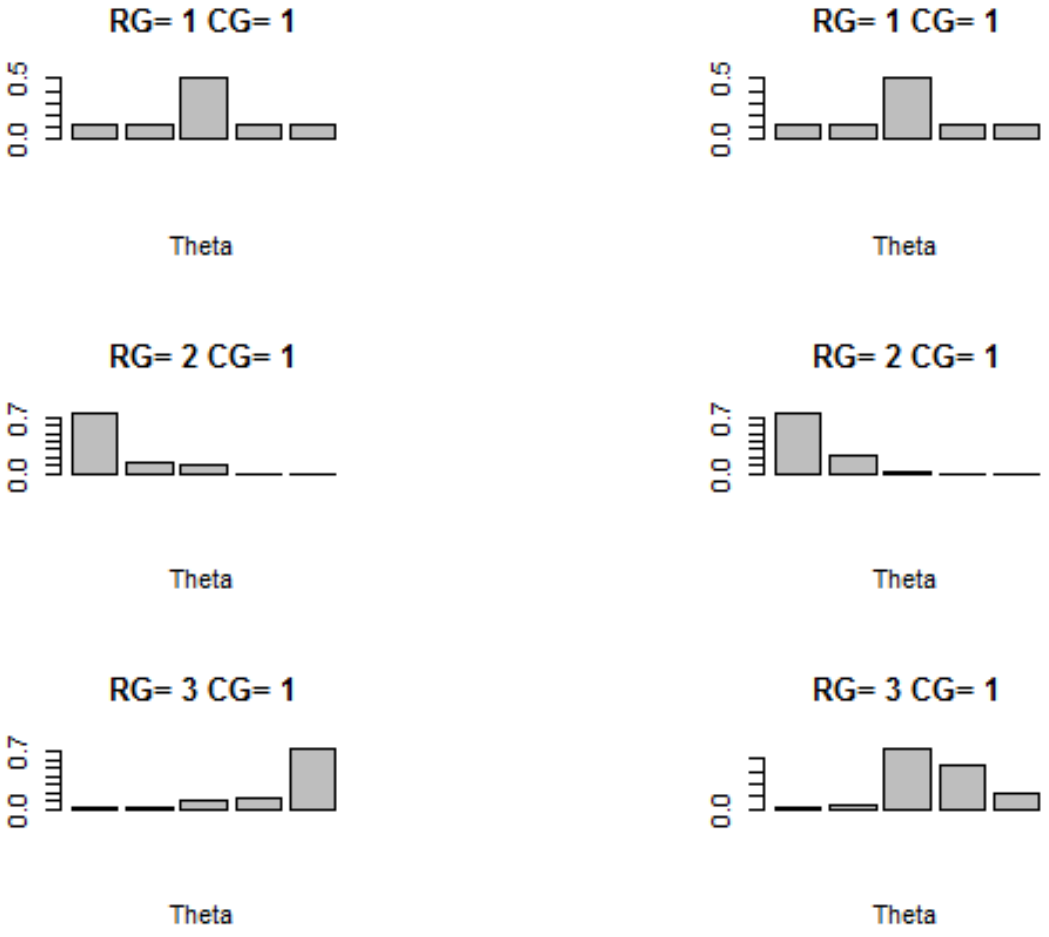


Figure 4: Probability distribution for θ_{rjk} for the POM and TOM

Assuming independence over the columns and independence over the rows conditional on the columns, the model's likelihood is

$$L(\phi, \kappa|\mathbf{Y}) = \prod_{j=1}^{p} \sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{ick}^{I(y_{ij}=k)} \tag{6}$$

The incomplete information likelihood for the column-cluster case has $v = (q-1) + 3(C-1)$ parameters.

3.1.3 Bi-clustering

In this case, 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 $\pi_1, \ldots, \pi_C$ 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, γ_r, β_c, and δ_c for the row and column-cluster effects, the linear predictor becomes:

$$\text{Logit}[P(y_{ij} \leq k | i \in r \cap j \in c)] = \mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c)t_k \tag{7}$$

or alternatively

$$\theta_{rck} = \frac{\exp[\mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c)t_k]}{1 + \exp(\mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c)t_k)} - \frac{\exp(\mu_{k-1} - \alpha_r - \beta_c - (\gamma_r + \delta_c)t_{k-1})}{1 + \exp(\mu_{k-1} - -\alpha_r - \beta_c - (\gamma_r + \delta_c)t_{k-1})} \tag{8}$$

Where $\alpha_1 = \gamma_1 = \beta_1 = \delta_1 = 0$. Let $t_k = k - 1$ as before, the constraint to make sure the cumulative probabilities are non-decreasing is

$$\mu_k - \mu_{k-1} \geq (\gamma_r + \delta_c) \qquad \forall r, c, k \tag{9}$$

In this case, the likelihood for the TOM 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

$$L(\phi, \pi, \kappa | \mathbf{Y}) = \sum_{c_1=1}^{C} \cdots \sum_{c_p=1}^{C} \kappa_{c_1} \ldots \kappa_{c_p} \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rc_jk}^{I(y_{ij}=k)} \tag{10}$$

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

$$L(\phi, \pi, \kappa | \mathbf{Y}) = \sum_{r_1=1}^{R} \cdots \sum_{r_n=1}^{R} \pi_{r_1} \ldots \pi_{r_n} \prod_{j=1}^{p} \sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} \theta_{r_ick}^{I(y_{ij}=k)} \tag{11}$$

Likewise the former, due to requiring consideration of all possible allocations of the n rows to the R groups this expression is very expensive to compute. In either specification, equations 10 or 11, the incomplete data likelihood for this case has $v = (q-1) + 3(R + C - 2)$ parameters.

3.1.4 Estimation

In this section we use a Frequentist framework to maximise the likehoods described above and estimate the model parameters. In particular we use the EM algorithm (Dempster et al., 1977). In essence, the EM algorithm turns the problem into a missing data problem and then estimates the parameters using an iterative two-fold approach. It first estimates the missing data given some initial values for the parameters (E-step). Next, the parameters are estimated by maximizing the complete data likelihood given the estimated missing data (M-step). These new parameters in turn are feed to be E-step again and the process repeats until the paramaters converge, that is the change in the parameters is tiny, e.g. a relative change between iterations of say less than $1x10^{-6}$.

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 i unkwown and thus is regarded as missing data. Note that $\sum_{r=1}^{R} z_{ir} = 1$, and we define a (n, R) matrix $\mathbf{Z}$ as a membership matrix to gather together all the z_{ir}'s.

The complete data log-likelihood is

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

Given a value for the number of mixture components R, the EM algorithm proceeds as follows.

E-step: Update $\mathbf{Z}$. Given $\mathbf{Y}$ and values for π_r, μ_k, α_r and γ_r, estimate the expected values of z_{ir} and π_r as

$$E[z_{ir}|\mathbf{Y}, \pi_r, \mu_k, \alpha_r, \gamma_r] = \hat{z}_{ir} = \frac{\pi_r \prod_{j=1}^{p}\prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)}}{\sum_{a=1}^{R} \pi_a \prod_{j=1}^{p}\prod_{k=1}^{q} \theta_{ajk}^{I(y_{ij}=k)}} \tag{12}$$

and $E[\pi_r|\mathbf{Z}] = \hat{\pi}_r = \sum_{i=1}^{n} \hat{z}_{ir}/n$.

M-step: Numerically maximise the complete data log-likelihood. Given $\hat{z}_{ir}$ and $\hat{\pi}_r$ from the E-step maximise ℓ_c to obtain new values for the parameters μ_k, α_r, and γ_r

A new cycle starts when the parameters from the M-step are used in the E-step. This process repeats until estimates have converged. Note that θ_{rjk} is estimated using equation 2. Importantly, there is a risk of convergence to local maxima due to multimodality on the likelihood surface and thus it is important to start the EM algorithm with several initial values.

Column-clustering

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

E-step: Update $\mathbf{X}$. Given Y and initial values for κ_c, μ_k, β_c, and δ_c estimate the expected values of x_{jc} and κ_c as

$$E[x_{jc}|\mathbf{Y}, \kappa_c, \mu_k, \beta_c, \delta_c] = \hat{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n}\prod_{k=1}^{q} \theta_{ick}^{I(y_{ij}=k)}}{\sum_{a=1}^{C} \kappa_a \prod_{i=1}^{n}\prod_{k=1}^{q} \theta_{iak}^{I(y_{ij}=k)}} \tag{13}$$

and $E[\kappa_c|\mathbf{X}] = \hat{\kappa}_c = \sum_{j=1}^{p} \hat{x}_{jc}/p$.

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

$$\ell_c = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{c=1}^{C}\sum_{k=1}^{q} \hat{x}_{jc} I(y_{ij} = k) \log(\theta_{ick}) + \sum_{j=1}^{p}\sum_{c=1}^{C} \hat{x}_{jc} \log(\hat{\kappa}_c) \tag{14}$$

A new cycle starts when the parameters from the M-step are use in the E-step. This process repeats until the change in any of the parameters is smaller than 10^{-6}.

Bi-clustering

Let z_{ir} and x_{jc} be the latent row and column cluster memberships for each cell (i, j). As before $\mathbf{Z}$ and $\mathbf{X}$ 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 $(\mu, \pi, \alpha, \gamma, \kappa, \beta, \delta)$, we also incorporate the variational approximation employed by Govaert & Nadif (2005)

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

That is, conditonal on the ordinal response and the parameters the effect of the row and column clusters are independent. Given this approximation, 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: Update $\mathbf{Z}$ and $\mathbf{X}$. Given $\mathbf{Y}$ and initial values for ϕ estimate the expected values of z_{ir}, x_{jc}, π_r, and κ_c as

$$E[z_{ir}|\mathbf{Y}, \phi] = \hat{z}_{ir} = \frac{\pi_r \prod_{j=1}^{p} \left\{ \sum_{c=1}^{C} \kappa_c \prod_{k=1}^{q} \theta_{rck}^{I(y_{ij}=k)} \right\}}{\sum_{a=1}^{R} \pi_a \prod_{j=1}^{p} \left\{ \sum_{b=1}^{C} \kappa_b \prod_{k=1}^{q} \theta_{abk}^{I(y_{ij}=k)} \right\}} \tag{15}$$

$$E[x_{jc}|\mathbf{Y}, \phi] = \hat{x}_{jc} = \frac{\kappa_c \prod_{i=1}^{n} \left\{ \sum_{r=1}^{R} \pi_r \prod_{k=1}^{q} \theta_{rck}^{I(y_{ij}=k)} \right\}}{\sum_{b=1}^{C} \kappa_b \prod_{i=1}^{n} \left\{ \sum_{a=1}^{R} \pi_a \prod_{k=1}^{q} \theta_{abk}^{I(y_{ij}=k)} \right\}} \tag{16}$$

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

M-step: Numerically maximise the complete data log-likelihood. Given $\hat{x}_{ir}$ and $\hat{x}_{jc}$ from the E-step maximise ℓ_c to obtain new values for the parameters $\mu_k, \alpha_r, \gamma_r, \beta_c$, and δ_c

$$\ell_c = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{k=1}^{q} \hat{z}_{ir} \hat{x}_{jc} I(y_{ij} = k) \log(\theta_{rck}) + \sum_{i=1}^{n} \sum_{r=1}^{R} \hat{z}_{ir} \log(\hat{\pi}_r) + \sum_{j=1}^{p} \sum_{c=1}^{C} \hat{x}_{jc} \log(\hat{\kappa}_c) \tag{17}$$

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 for the bi-clustering (equation 8).

3.1.5 Model selection

Model selection criteria for finite mixtures is an active area of research in Statistics with no theoretical foundations completely developed yet to date. In this section, we use the integrated classification criterion (ICL) by Biernacki et al. (2000). In the future, one of our goals is to compare the performance of different model selection criteria for finite mixtures of repeated ordinal data, see section 5 for details.

Given that we are using cluster models based on finite mixtures and calculating likelihoods, we could in principle use likelihood-based model selection criteria, such as AIC and BIC, to compare amongst models. Simulation studies showed however that AIC and BIC tend to overestimate the number of mixture components (McLachlan & Peel, 2000; Cubaynes et al., 2012). More importantly though, comparing finite mixture models with different number of clusters violates classical regularity conditions (Cramér, 1946) and therefore the use of criteria such as AIC (Akaike, 1973) or BIC (Schwarz, 1978) is doubtful. In particular, ML estimates under the null hypothesis (reduced model) are on 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 always be re-express as a mixture of "g+1" components by doubling up one of its components and halving its mixture probability. 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 special cases but not for mixtures of ordinal data.

Although not exempt from this problem, 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. 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 doesn't require the evaluation of the incomplete information likelihood L. This evaluation is computationally very expensive, specially for large datasets and when dealing with bi-clustering. For bi-clustering for example, 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 (equations 10 and 11).

The ICL is formed from the maximised complete data likehood ℓ_c (for example equation 17 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 separatation 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

$$ICL - BIC = -2\ell_c(\hat{\phi}, \mathbf{Y}) + v\log(np) \tag{18}$$

Where ϕ and v are the set and number of model parameters, respectively. For instance in the case of row-clustering: $\phi = (\mu, \alpha, \gamma, \pi)$, $v = (q-1) + 3(R-1)$, n= number of rows, p = number of columns, q= number of ordinal outcomes, and R the number of row-clusters.

3.2 Repeated measures Proportional Odds Model with clustering

This subsection presents a model where the correlation between observations is explicitely modelled. We do so by augmenting the linear predictor of the POM with row-clustering with a time random effect that varies by cluster. We called this model the repeated measures POM with clustering.

3.2.1 Correlation structure: random effects by cluster and occasion

Row-clustering

In addition to the notation defined earlier in this section, let $\beta_{rj} \sim N(0, \sigma_r^2)$ be a random effect by cluster and occasion. The linear predictor for the repeated measures POM with row-clustering is then

$$\text{Logit}[P(y_{ij} \leq k | i \in r)] = \mu_k - \alpha_r - \beta_{rj} \tag{19}$$

This expression is the same at the one for the POM with row-clustering with the addition of the β_{rj}, which is a term allows the repeated observations within each cluster to come from the same distribution $N(0, \sigma_r^2)$ and therefore share a common variance σ_r^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 2.2, these integrals are usually solved 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 could be very expensive computationally. Here we take a different approach and estimate it using Bayesian methods.

3.2.2 Estimation

In a Bayesian setting, both data and parameters in the model are random and thus we need to specify distributions for each of them. In particular, in addition to the likelihood which specifies the distribution of the data conditional on the parameters we need to specify *prior* distributions for the parameters. We then obtain the distribution of the parameters given the data, e.g. the *posterior*, by using Bayes theorem. For instance, given a set of parameters ϕ, an response $\mathbf{Y}$, and prior $\omega(\phi|\mathbf{Y})$, the posterior is

$$\omega(\phi|\mathbf{Y}) = \frac{\ell(\phi|\mathbf{Y})\omega(\phi)}{\int \ell(\phi|\mathbf{Y})\omega(\phi)d\phi}$$

where $\ell(\phi|\mathbf{Y})$ the model's likelihood. Rather than computing $\omega(\phi|\mathbf{Y})$, we may instead draw samples from it.

In the case of the repeated measures POM with row-clustering, we use a Bayesian Hierarchical Model (BHM) and specify the likelihod and priors as follows:

$$\begin{aligned}
y_{ij} \mid \mu, \alpha_{r_i}, \beta_{r_i j}, r_i &\sim \text{POM}(\mu, \alpha_{r_i}, \beta_{r_i j}) && i = 1 \dots n,\ j = 1 \dots p,\ r = 2 \dots R \\
\mu'_k \mid \sigma^2_\mu &\overset{iid}{\sim} \text{Normal}(0, \sigma^2_\mu) && \mu_k = \mu'_{(k)},\ \text{if } \mu_k > \mu_{k-1},\ k = 1 \dots (q-1),\ \mu_q = \infty \\
\alpha_r \mid \sigma^2_\alpha &\sim \text{Normal}(0, \sigma^2_\alpha) && r = 2 \dots R,\ \alpha_1 = 0 \\
\beta_{rj} &\sim \text{Normal}(0, \sigma^2_r) && r = 2 \dots R,\ j = 1 \dots p,\ \beta_{1j} = 0\ \forall j \\
r_i \mid \pi &\sim \text{Categorical}(\pi) && i = 1 \dots n \\
\sigma^2_\mu &\sim \text{Inverse Gamma}(a_\mu, b_\mu) && \\
\sigma^2_\alpha &\sim \text{Inverse Gamma}(a_\alpha, b_\alpha) && \\
\sigma^2_r &\sim \text{Inverse Gamma}(a_r, b_r) && r = 2 \dots R \\
\pi &\sim \text{Dirichlet}(\psi) &&
\end{aligned}$$

With (fixed known) hyperparameters: $\psi, a_\alpha, b_\alpha, a_\mu, b_\mu, a_r$, and b_r.

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^2_r, 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, e.g they have the same probability to have an outcome k given that they all are in cluster r. This also allows the model to be parsimonous because we avoid needing to use $n-1$ terms for individuals and instead only use $R-1$ in the linear predictor. Finally, the occasions level is where the random effects by cluster and occasion is realised β_{rj} coming from a $N(0, \sigma^2_r)$. Figure 5 shows a graphical representation of the model.

It is important to note that when working with BHM the number of mixture components, the number of row groups R in the repeated measures POM with row-clustering, is fixed and exogenous to the model. Relaxing this assumption is a future natural extension and is detailed in section 5.

Although not implemented fully yet, we will use a Markov-Chain Monte-Carlo (MCMC) sampling scheme that nests Gibbs and Metropolis-Hastings steps. This nested sampling scheme is necesary because the full conditional distributions are non-standard. The model will be implemented in R and BUGS.

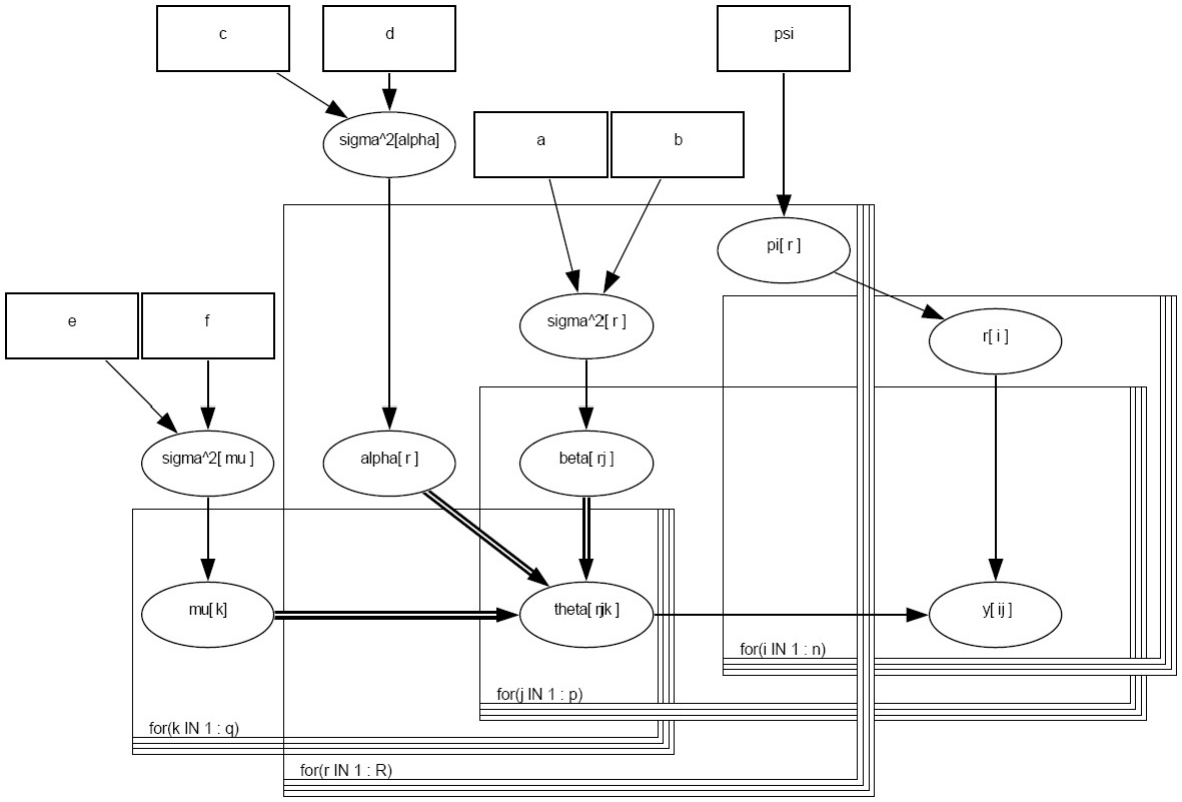


Figure 5: Graphical representation of repeated measures POM with row-clustering

3.2.3 Model selection

We use the Deviance Information Criterion (DIC) for model selection. Spiegelhalter et al. (1998) showed that it is an asymptotic generalization of the AIC in the context of hierchical models. It has two-parts: a classical estimate of fit $\ell(\hat{\phi}, \mathbf{Y})$ and twice the estimated number of model parameters v.

$$DIC = \ell(\hat{\phi}, \mathbf{Y}) + 2v \tag{20}$$

In the case of repeated measures POM with row-clustering, $\phi = (\mu, \alpha, \sigma^2, \pi)$ and $v = (q-1) + 3(R-1)$. A future extension, will use Bayes Factors and Reversible Jump Markov Chain Monte Carlo (RJMCMC) for model selection. See section 5 for details.

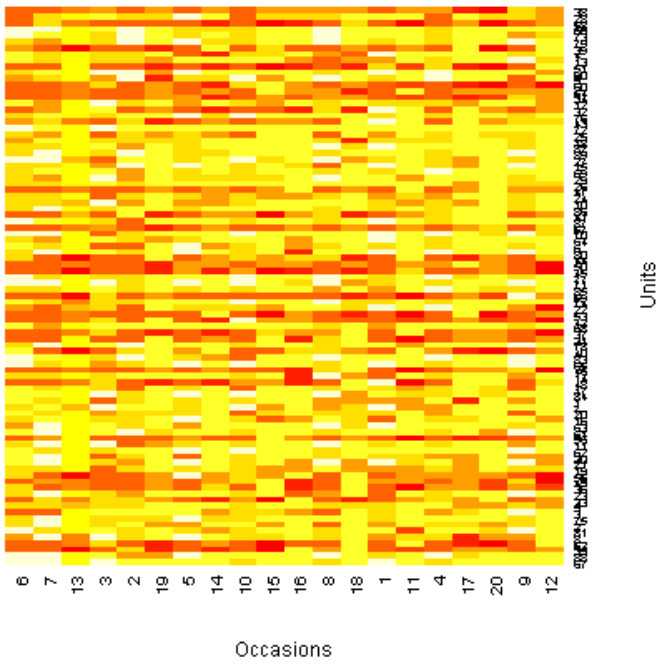
3.3 Visualisation

One of the goals of our research is to explore new methods of displaying ordinal data. This includes the data itself but also the parameters and predictions of fitted models. In this section, we use heatmaps to visualise ordinal data and the resulting cluster from our models. Heatmaps are a graphical representation of a matrix where cell values are represented using different colors. In our case, they represent the ordinal outcomes with darker colors indicating a higher ordinal outcome.

Figure 6 shows heatmaps for simulated data from a bi-clustered POM ($R = 3$ and $C = 2$) with five ordinal outcomes ($q = 5$) where rows ($n = 90$) and columns ($p = 20$) represent observations and occassions for each observation. [1]

[1] We used the following parameters: $\alpha = (0, -3, 2)$, $\gamma = (0, 0.25, -0.25)$, $\beta = c(0, 2)$, $\delta = c(0, 0.2)$, and

Data prior to clustering



Bi-clustered data (POM,3 row and 2 column groups)

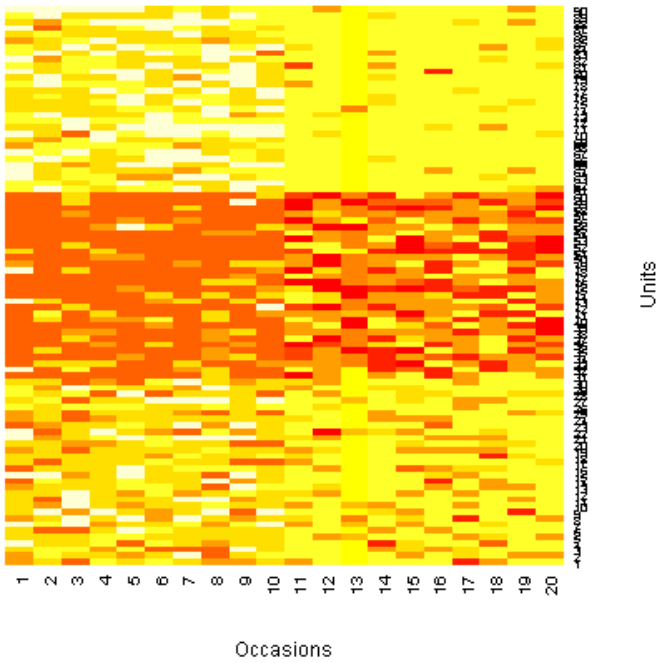


Figure 6: Heatmaps for simulated data

The first category corresponds to white cells, the intermediate ones to yellow and the higher one to red cells. The top heatmap shows the data prior to clustering and the bottom one the clustered data with three row and two column groups. Although both represent the same data, no pattern is evident in the former whereas is easy to see the biclustered pattern in the latter one.

$\mu = (-1.95, -1.10, 1.10, 1.95)$ The mixture probabilities are $\pi = (0.33, 0.33, 0.33)$ and $\kappa = (0.5, 0.5)$.

It is important to mention that heatmaps should only be used to visualise the best fitting model, after statistical estimation of several models, and not to compare among models since the human eye tends to see patterns in any given image (Wilkinson & Friendly, 2009).

4 Applications: case study and simulations

To illustrate the TOM with clustering, we perform simulations and also apply it to real data from the Household, Income and Labour Dynamics in Australia (HILDA).

Case study

We use 2001-2011 self-reported health status (SRHS) from HILDA [2]. HILDA is a household-based panel study which began in 2001 that collects information about economic and subjective well-being, labour market dynamics and family dynamics. The wave 1 panel consisted of 7,682 households and 19,914 individuals.

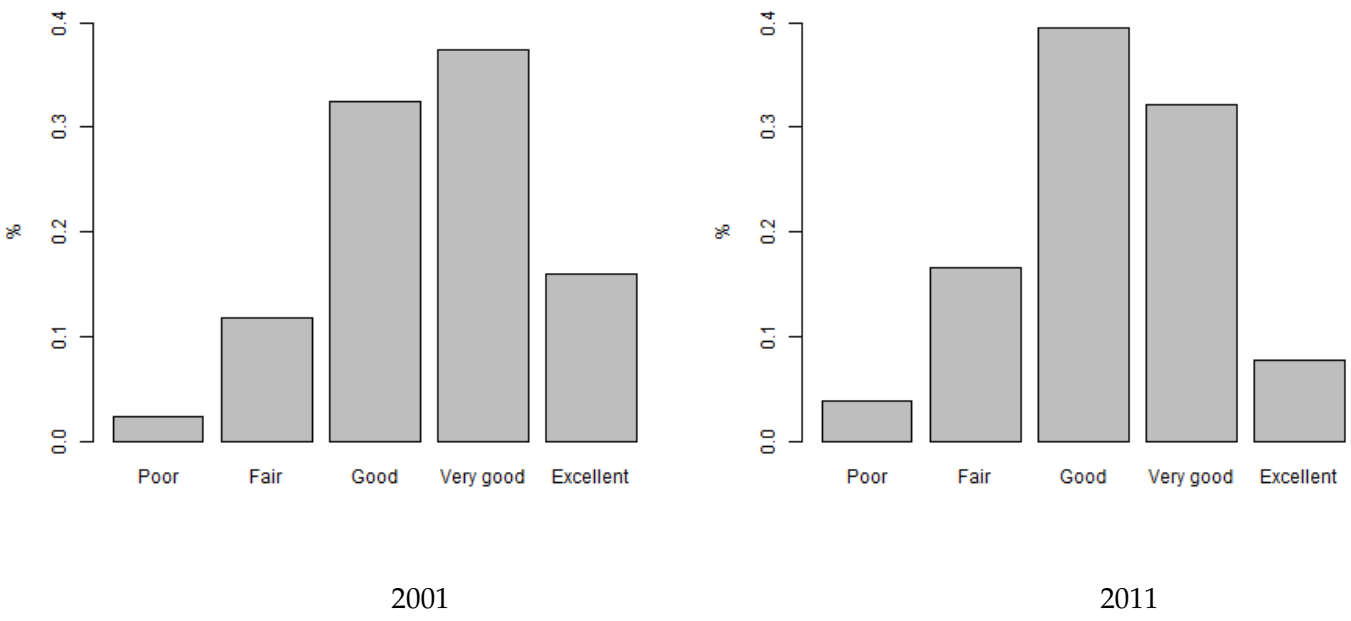


Figure 7: Distribution of Self-Reported Health Status (SRHS) in 2001 and 2011 in HILDA

SRHS is an ordinal variable with 5 categories: poor, fair, good, very good and excellent. [3]. We use individuals with complete records over 2001 to 2011, that is we have 11 occasions of SHRS from the same individuals.

Figure 7 shows the distribution of SRHS in 2001 and 2011. In 2001, most individuals reported "Very Good" and "Good" health. About an eight reported their health as "Excellent" and about a tenth as "fair". A very low number of individuals said their health was "poor".

[2]This document uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. 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). The findings and views reported here, however, are those of the author and should not be attributed to either DSS or the Melbourne Institute.

[3]Every year individuals were asked : "In general, would you say your health is:" and they chose between these 5 ordinal categories

In contrast to that, in 2011 the same individuals reported lower health levels. "Excellent" and "Very Good" answers decreased and "Poor" and "Fair" increased. SRHS's distribution shifted to the left and is more symetric in 2011 than in 2001.

Furthermore, for each individual SRHS is highly correlated across time . Table 1 presents the 2001-2011 transitions between ordinal categories for all individuals. Diagonal proportions are very high, about 40%, and 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 1: 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 estimate the parameters of the TOM with clustering, we use a sample of 136 individuals that had their SRHS recorded in all the waves. Therefore, we estimate it from a sample of 136 rows (n), 11 columns (p) and 5 ordinal categories (q). In addition to that, we also fit POM with clustering and models with row and columns effects. Model comparison is carried out using the ICL, as explained in section 3.1.5, and also the AIC and BIC.

Table 2 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. The model with the best fit is the TOM with five row clusters with a ICL-BIC of 2922. It has a total of 16 parameters ($\mu_k, \alpha_r, \gamma_r$ and π_r where $k = 1 \dots 4$ and $r = 1 \dots 4$). 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 present here.

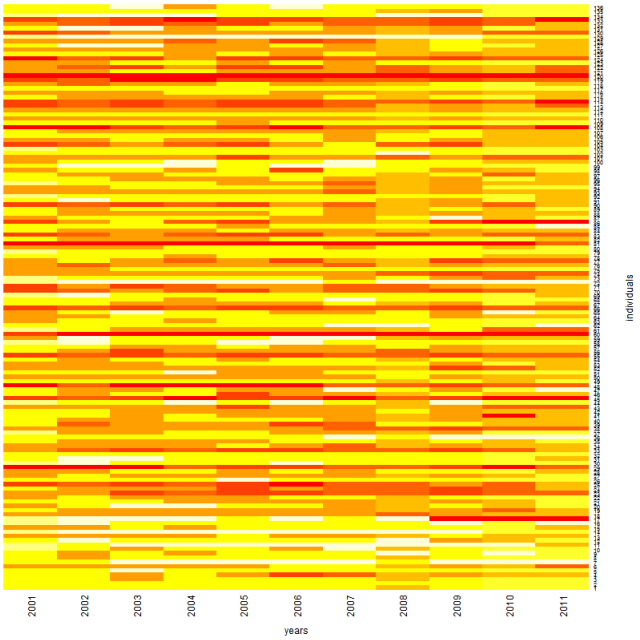
In sum, results for this subsample of the 2001-2001 SRHS suggest that the individuals could be grouped into five clusters.

Table 2: 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 fixed effects	$\mu_k - \alpha_i$	136	1	139	2446	4200	
Col fixed effects	$\mu_k - \beta_j$	1	11	14	4094	4271	
Row and Col fixed 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 col clustering	$\mu_k - \beta_c$	1	2	6	4092	4124	4134
POM biclustering	$\mu_k - \alpha_r - \beta_c$	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 col clustering	$\mu_k - \beta_c - \delta_c t_k$	1	2	7	4095	4132	4132
TOM biclustering	$\mu_k - \alpha_r - \beta_c - (\gamma_r + \delta_c) t_k$	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

How 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 below. Individuals and occasions are shown in rows and columns and cell colors represent ordinal categories . As shown in Figure 8, 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.

Unordered data



TOM row-clustered data (5 groups)

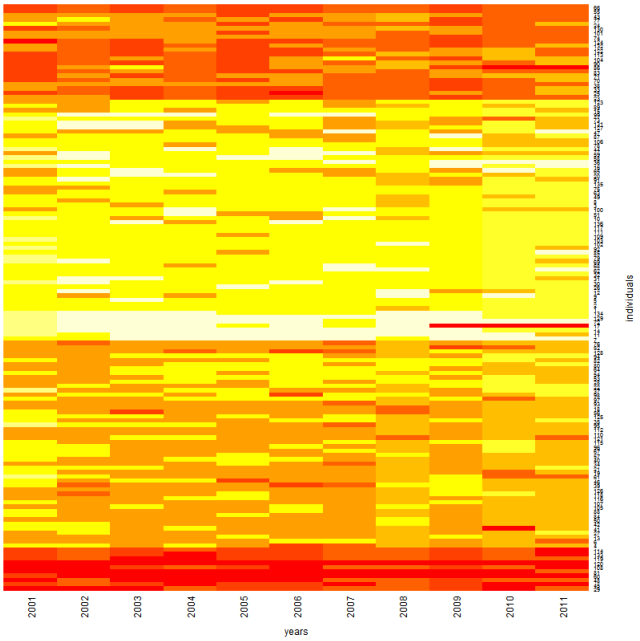


Figure 8: Heatmaps for Self-Reported Health Status in HILDA

4.1 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 the 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 $\pi = (0.33, 0.33, 0.33)$ and $\kappa = (0.5, 0.5)$.

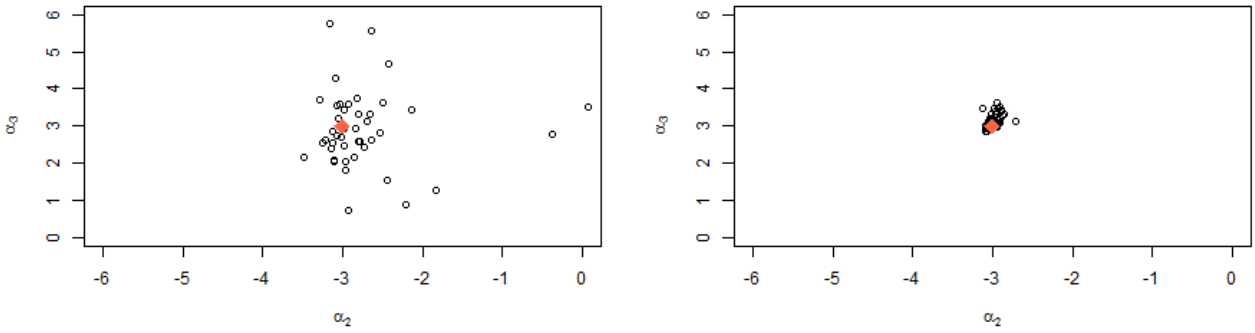
We finally set the number of ordinal categories equal to five $(q = 5)$, and estimate the TOM under four scenarios with increasing sample size, e.g. number of rows (n) and columns (p) increases. Scenario 1 presents $n = 60$ and $p = 10$, Scenario 2 presents $n = 120$ and $p = 10$, Scenario 3 presents $n = 300$ and $p = 10$, and Scenario 4 presents $n = 1200$ and $p = 10$. Table 3 shows the true model parameters, and the mean and standard errors (SE) of their estimations for 50 runs on each scenario.

Table 3: Simulation results for TOM biclustering with R=3, C=2

		n=60		n=120		n=300		n=1200	
Parameter	true	mean	se	mean	se	mean	se	mean	se
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	0.28	0.10	0.19	0.06	0.05	0.03	0.01	0.01
α_2	-3.00	-3.00	0.00	-3.00	0.00	-3.00	0.00	-3.00	0.00
α_3	3.00	4.33	0.63	3.32	0.28	3.26	0.06	3.16	0.03
γ_1	0.00	-0.09	0.04	-0.08	0.05	-0.01	0.01	0.00	0.00
γ_2	-0.20	0.18	0.03	0.21	0.02	0.18	0.01	0.20	0.00
γ_3	0.50	1.74	0.60	0.63	0.03	0.61	0.02	0.68	0.03
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

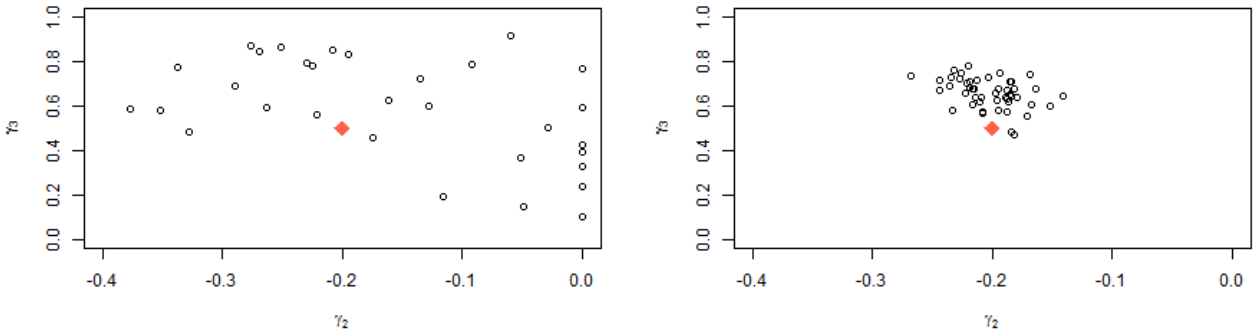
As it can be seen, most means get closer to their true values as the number of rows in the sample (n) increases from Scenario 1 to 4. Further, standard errors also decrease with higher n. For example, in the case of $\beta_2 = 2$ the mean of the estimates goes from 1.85 with a SE of 0.05 in scenario 1 to 2.01 with a SE of 0.01 in Scenario. The one exception to the above, is the case of γ_3 which is overestimated. Its true value is 0.5 but the mean of the estimates ranges from 1.74 to 0.60 in the different scenarios, being 0.68 in Scenario 4. Importantly, however, this doesn't affect the estimates for the mixture proportions $\pi = (0.33, 0.33, 0.33)$ and $\kappa = (0.5, 0.5)$. The mean of the estimates on each scenario are very close to the true proportions even with small n.

As a further illustration, Figures 9 to 10 show the estimated parameters for all 50 simulations scenarios 1 and 4 for α, γ, β, and δ. 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.



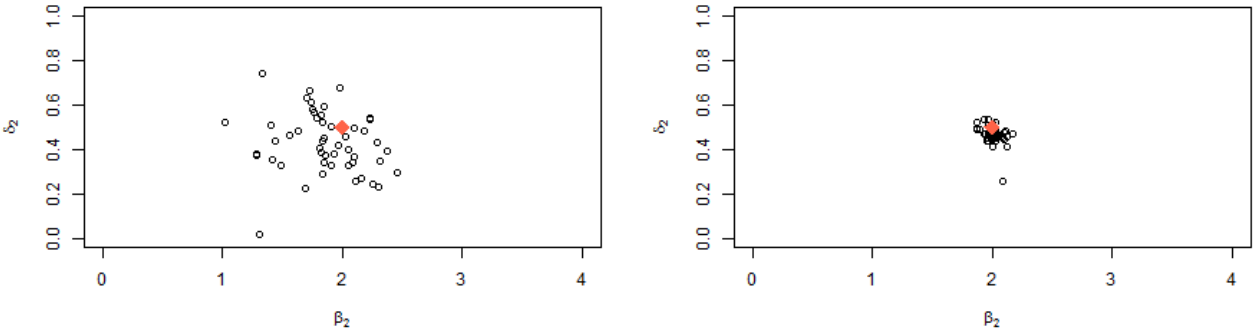
n=60 n=1200

Figure 9: Simulation results for Biclustering $\alpha_2 = -3$ and $\alpha_3 = 3$



n=60 n=1200

Figure 10: Simulation results for Biclustering $\gamma_2 = -0.2$ and $\gamma_3 = 0.5$



n=60 n=1200

Figure 11: Simulation results for Biclustering $\beta_2 = 2$ and $\delta_2 = 0.5$

5 Research Goals

So far, this proposal has reviewed the relevant literature (section 2) and developed two cluster models based on finite mixtures for repeated ordinal data, the TOM and POM with clustering (section 3) . We also have estimated the former using simulated and real data, including visualising the results through heatmaps (section 4). In the next two years we have the following research goals (RG).

- RG1: To determine the impact of different correlation structures on estimated clusters for repeated ordinal data
- RG2: to compare the performance of different model selection criteria using a Bayesian framework
- RG3: To explore new visualisation methods for data, parameters and predictions
- RG4: To jointly model repeated ordinal and continuous data

5.1 To determine the impact of correlation structures on estimated clusters for repeated ordinal data

We will develop a range of models here to assess the impact of correlation structures on estimated clusters. This includes models for two-way (unit by time period) and three-way (unit by question by time period) data. With regard to the former, in addition to the repeated measures POM with clustering (section 3.2) we will incorporate the correlation due to repeated measurements as an autoregressive order of order one (AR1) and also bi-cluster individuals and time periods aiming to unfold latent temporal states (block correlation structure). In the case of three-way data, the scope is large but we will build on the above models to see what is possible. For instance, we could pose an association structure where there is bi-clustering in individuals and questions and an AR1 temporal correlation. Finally, in order to showcase develop models we will applied models to datasets with different correlation structures. Amongst others, we propose applied it to : 2007-2011 income survey data (extremely poor, poor, medium, high and very high) from the Peruvian National Survey of Households (ENAHO/Peru) , 2006-2012 household work status response in NZ (none member working, female working, male working, both working), and crimes severity overtime (data to be obtained).

5.2 To compare the performance of different model selection criteria for finite mixtures of repeated ordinal data

Using a Bayesian framework, we will extent the work of Arnold et al. (2010) for the binary case and compare the performance of DIC (section 3.2), Bayes Factors and Reversible Jump Markov Chain Montecarlo (RJMCMC) for the repeated measures POM with clustering. Would more complex approaches as model averaging using RJMCMC be better for instance? An important caveat of the Bayesian Hierarchical model (BHM) presented in section 3.2, is that it is conditional on the unknown number of mixture components. That is, the model holds if the number of mixture components is the right one. In some cases, models with different number of components are very similar so which provides the best fit is not clear (multimodal posterior). In these cases, more complex approaches such as RJMCMC could prove useful as they could incorporate this unknown number of mixture components as parameter to be estimated and provide a model average if necessary. On the other hand, model comparison using frequentist information criteria will also be performed. Here we

will compare the AIC, BIC, ICL, and ICL-BIC for finite mixtures of repeated ordinal data. We will build on the work of McLachlan & Peel (2000) and Brame et al. (2006) for the normal and poisson mixtures. Simulation studies will be used to compare the models in terms of bias, precision and quality of classification (correctly, assigning individuals into the mixture components).

5.3 To explore new visualisation methods for data, parameters and predictions

We have used heatmaps up until now, but are there any other visualisation methods that could provide an additional visual representation of repeated ordinal data? We will explore the possibilty of extending mosaic plots to the repeated measures case but also triallying transition probability color plots, also called decile transition matrices, that show the initial and final distribution of the ordinal response.

5.4 To jointly model repeated ordinal and continuous data

Joint modelling of ordinal responses and other kind of variable are a further extension to the methods proposed here. This includes clustering of repeated ordinal data and continuous data collected over the same individuals. For instance, household work status overtime (none member working, female working, male working, both working) from RG1 could be jointly modelled with household income (continuous) in order to identify potential patterns in a latent variable that could be causing both. This problem is of interest in Economics as current approaches analyse them separately or use other estimation methods, e.g. simulated methods of moments (McFadden, 1989; Card & Hyslop, 2005). We aim to to answer this question through proposed joint models. Professor Dean Hyslop (SEF, VUW) has expressed interest applying these methods.

6 Plan

6.1 PhD thesis structure

A tentative structure for this PhD thesis, *Cluster analysis of repeated ordinal data: A model approach based on finite mixtures*, and the associated RGs is presented below:

1. Introduction
2. Literature review
3. No correlation structure: The trend odds model - RG1
4. A simple correlation structure: The repeated measures proportional odds model - RG1
5. More complex correlation structures for two-way data - RG1
6. Other correlation structures for three-way data - RG1
7. Model selection - RG2
8. Visualisation tools - RG3
9. Case studies - RG1-R4
10. Joint modelling of repeated ordinal and continuous data - RG4
11. Conclusions - RG1- RG4

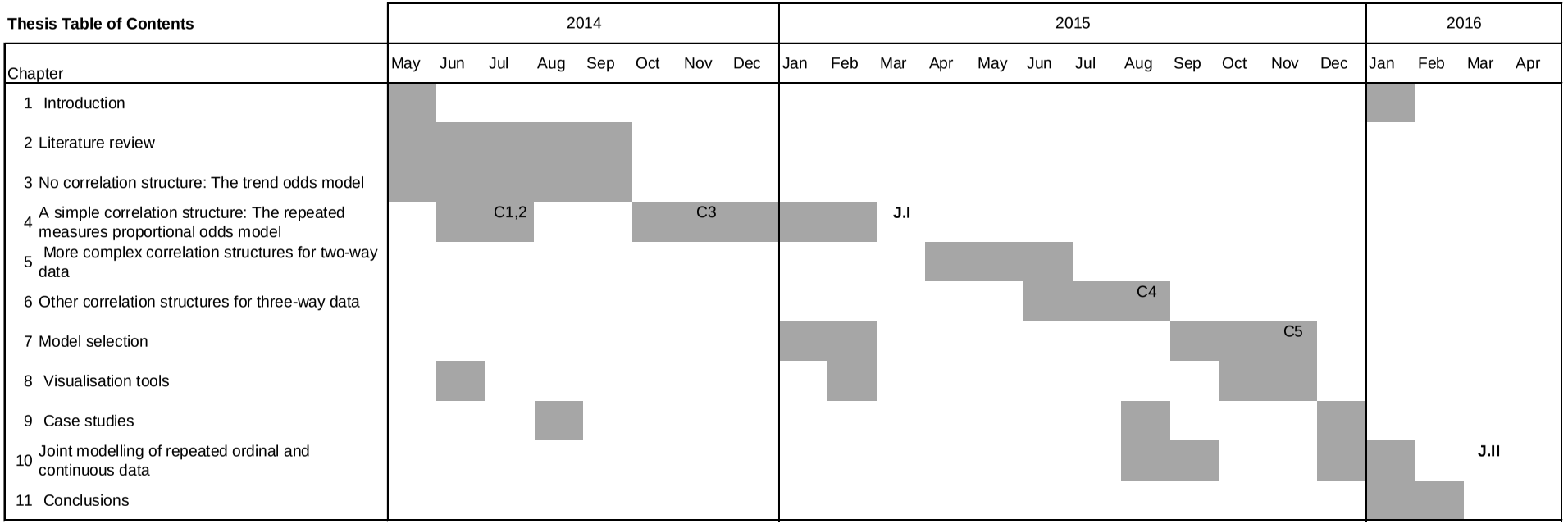
6.2 Timeline

In addition to these chapters, outputs include (2) publications and assistance to (5) local and international conferences to present the research. The proposed journal articles are:

1. Cluster analysis for repeated ordinal data: A model-based approach using finite mixtures. Chapters 2-9
2. Modelling the dynamics of wage and employment of families NZ using joint models for ordinal and continuous data, chapter 10

The workload for next 2 years is detailed in the next page. It includes publications and conferences indicative dates.

PhD Plan: thesis, publications and conferences



Chapter ready for submission

J Journal article ready for submission (2)
- I. Chaps 2-4, 7-9
- II. Chap 5-10

C Conference
1. ISBA, Mexico Cancun
2. NZAE, Auckland
3. NZSA, Wellington
4. JSM ASA, Seattle, USA
5. NZAE, NZ

References

Agresti, A. (2010). *Analysis of ordinal categorical data, 2nd edition*. Wiley Series in Probability and Statistics.

Agresti, A. (2013). *Categorical data analysis, 3rd edition*. John Wiley & Sons.

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* (p. 267-281).

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

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

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

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.

Bock, R. D., & Jones, J. (1968). The measurement and prediction of judgment and choice.

Brame, R., Nagin, D. S., & Wasserman, L. (2006). Exploring some analytical characteristics of finite mixture models. *Journal of Quantitative Criminology*, *22*(1), 31–59.

Capuano, A., & Dawson, J. (2012). The trend odds model for ordinal data. *Statistics in Medicine*, *32*, 22502261.

Card, D., & Hyslop, D. R. (2005). Estimating the effects of a time-limited earnings subsidy for welfare-leavers. *Econometrica*, *73*(6), 1723–1770.

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.

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.

Everitt, B., Landau, S., & Leese, M. (2001). Cluster analysis. 2001. *Arnold, London*.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models, 2nd edition*. Springer New York.

Fernandez, F., Arnold, R., & Pledger, S. (2014). Fuzzy clustering for the ordered stereotype model via finite mixtures. *Computational Statistics and Data Analysis (submitted)*.

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.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2) (No. 1). Springer.

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

Kedem, B., & Fokianos, K. (2002). *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.

Labiod, L., & Nadif, M. (2011). Co-clustering for binary and categorical data with maximum modularity. In *Icdm* (pp. 1140–1145).

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.

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

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), (chap. The Sage handbook of quantitative methodology for the social sciences). SAGE Publications.

Manly, B. F. (2005). *Multivariate statistical methods: a primer*. CRC Press.

Matechou, E., Liu, I., Farias, M., & Gjelsvik, B. (2014). Biclustering models for ordinal data. *Psycometrika (submitted)*.

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.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, *57*(5), 995–1026.

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

Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, *4*, 1-274.

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

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.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.

Snell, E. (1964). A scaling procedure for ordered categorical data. *Biometrics*, 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., Carlin, B. P., & Van der Linde, A. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models* (Tech. Rep.). Research Report, 98-009.

Thompson, R., & Baker, R. (1981). Composite link functions in generalized linear models. *Applied Statistics*, 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 & Data Analysis*, *22*(5), 537–557.

Vermunt, J. K., & Hagenaars, J. A. (2004). Methods in human growth research. In R. Hauspie, N. Cameron, & L. Molinari (Eds.), (chap. Ordinal longitudinal data analysis). Cambridge University Press.

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.

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