Computational Statistics

Bayesian model-based clustering for longitudinal ordinal data

--Manuscript Draft--

Manuscript Number:	COST-D-18-00104
Full Title:	Bayesian model-based clustering for longitudinal ordinal data
Article Type:	Original Paper
Keywords:	Classification; Latent transitional models; Correlated Data; Finite mixture models; Widely Applicable Information Criterion (WAIC); MCMC

Corresponding Author:	Roy Costilla, PhD University of Queensland AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Queensland
Corresponding Author's Secondary Institution:	
First Author:	Roy Costilla, PhD
First Author Secondary Information:	
Order of Authors:	Roy Costilla, PhD
	Ivy Liu
	Arnold Richard
	Daniel Fernandez
Order of Authors Secondary Information:	

Funding Information:	Marsden Fund (16-VUW-062)	Dr Ivy Liu

Abstract:	Traditional cluster analysis methods used in ordinal data, for instance k-means and hierarchical clustering, are mostly heuristic and lack statistical inference tools to compare among competing models. To address this we propose a latent transitional model, a finite mixture model that includes both observed and latent covariates and apply it for the first time to the case of longitudinal ordinal data. This model-based clustering model is an extension of the Proportional Odds model and includes a first-order transitional term, occasion effects and interactions which provide flexible ways to capture different time patterns by cluster as well as time-heterogeneous transitions. We estimate model parameters within a Bayesian setting using a Markov chain Monte Carlo (MCMC) scheme and block-wise Metropolis-Hastings sampling. We illustrate the model using 2001-2011 self-reported health status (SRHS) from the Household, Income and Labour Dynamics in Australia (HILDA) survey. SRHS is recorded as an ordinal variable with five levels: poor, fair, good, very good and excellent. Using the Widely Applicable Information Criterion (WAIC) for model comparison, we find evidence for six latent groups. Transitions in the original data and the estimated groups are visualized using heatmaps.

Computational Statistics manuscript No.
(will be inserted by the editor)

Bayesian model-based clustering for longitudinal ordinal data

Roy Costilla · **Ivy Liu** · **Richard Arnold** · **Daniel Fernández**

Received: date / Accepted: date

Abstract Traditional cluster analysis methods used in ordinal data, for instance k-means and hierarchical clustering, are mostly heuristic and lack statistical inference tools to compare among competing models. To address this we propose a latent transitional model, a finite mixture model that includes both observed and latent covariates and apply it for the first time to the case of longitudinal ordinal data. This model-based clustering model is an extension of the Proportional Odds model and includes a first-order transitional term, occasion effects and interactions which provide flexible ways to capture different time patterns by cluster as well as time-heterogeneous transitions. We estimate model parameters within a Bayesian setting using a Markov chain Monte Carlo (MCMC) scheme and block-wise Metropolis-Hastings sampling. We illustrate the model using 2001-2011 self-reported health status (SRHS) from the Household, Income and Labour Dynamics in Australia (HILDA) survey. SRHS is recorded as an ordinal variable with five levels: poor, fair, good, very good and excellent. Using the Widely Applicable Information Criterion (WAIC) for model comparison, we find evidence for six latent groups. Transitions in the original data and the estimated groups are visualized using heatmaps.

Keywords Classification · Latent transitional models · Correlated Data · Finite mixture models · MCMC · Widely Applicable Information Criterion (WAIC)

R. Costilla (✉ r.costilla@imb.uq.edu.au)
Institute for Molecular Bioscience. Queensland Bioscience Precinct (Building 80). The University of Queensland. 306 Carmody Road. St Lucia Qld 4072 Australia.

R. Arnold and I. Liu
School of Mathematics and Statistics, Victoria University of Wellington, New Zealand.

D. Fernández
Institut de Recerca Sant Joan de Déu, Parc Sanitari Sant Joan de Déu, CIBERSAM, Spain.
School of Mathematics and Statistics, Victoria University of Wellington, New Zealand.

1 Introduction

A variable with an ordered categorical scale is called ordinal Agresti (2013). That is, ordinal data are categorical data where the outcome categories have a logical order and thus the order of the categories matters. In his seminal paper, Stevens (1946), called a scale ordinal if "any order-preserving transformation will leave the scale form invariant" (p. 679). Examples of ordinal responses are: socio-economic status (low, medium, high), disease severity (not infected, initial, medium, advanced), agreement with a given statement (strongly disagree, disagree, neutral, agree, strongly agree) and other variables that use the Likert and Braun-Blanquet scales.

Ordinal data are very common, but they might be treated wrongly in several ways. For instance, it is common to assign numerical scores to ordinal categories, often equally spaced scores, which might be an incorrect and restrictive assumption. Moreover, treating the ordinal responses as if they were continuous could lead to predicted values outside the range of possible ordinal outcomes, and could produce misleading results due to "floor" and "ceiling" effects on the dependent variable (see (Agresti, 2010, Section 1.3.1)). By assigning numerical scores, traditional cluster approaches such as hierarchical clustering (Kaufman and Rousseeuw, 1990), association analysis (Manly, 2005), and partition optimization methods like k-means clustering (MacQueen, 1967) may apply. However, these methods are not based on likelihoods and thus statistical inference tools are not available and model selection criteria can not be used to compare different models. 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.

In the literature, ordinal data are often analysed by modelling the cumulative probabilities of the ordinal response and using a link function, usually logit or probit. The Proportional Odds Model (POM) by McCullagh (1980) is a cumulative logit model and is the most popular model to analyse ordinal data. The Proportional Odds property that gives the model its name 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. Liu and Agresti (2005) and Agresti (2010) described various proportional odds version models using adjacent-categories logits, cumulative logits (McCullagh, 1980), and continuation-ratio logits (McCullagh and Nelder, 1989).

Further challenges are posed with repeated measurements of an ordinal response, such as in longitudinal studies. For these two-way data (unit and time period), the correlation structure among repeated measures also needs to be accounted for. Diggle et al (2002) and Agresti (2013) discussed three main approaches to the analysis of such data: marginal models, subject-specific models, and transitional models. Transitional models include past responses as predictors, that is; they model the current response conditional on past responses and potentially other explanatory variables. A very popular transitional model is the first-order Markov model in which the current response is assumed to depend only on the immediately preceding response (Diggle et al, 2002; Kedem and Fokianos, 2005; Agresti, 2013).

When transition models also include latent variables they are known as Markov transition, latent transition and mixture-of-experts Markov models. Latent transition

models have been used for model based clustering of longitudinal data and time series of continuous and categorical nature (Frydman, 2005; Pamminger et al, 2010; Frühwirth-Schnatter et al, 2012; Cheon et al, 2014). For instance, Pamminger et al (2010) and Frühwirth-Schnatter et al (2012) presented a mixture-of-experts time-homogeneous Markov models to cluster categorical time series which also allowed for covariates. Models were estimated within a Bayesian approach, compared using several information criteria and illustrated using wage and income mobility in Austria. On the other hand, Frydman (2005) and Cheon et al (2014) developed more restricted versions of the 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 and known in advance. Frydman (2005) considers another constrained model where a transition matrix is estimated for a baseline cluster and the remaining ones are only scaled versions.

Model-based clustering approaches using finite mixtures have been proposed by several authors (McLachlan and Peel, 2000; Everitt et al, 2001), which mostly focus on either continuous, discrete or nominal responses, see literature reviews by Fraley and Raftery (2002), Marin et al (2005), and Melnykov and Maitra (2010). Finite mixture models allow the estimation of both latent group effects and memberships and are often fitted using the Expected-Maximisation (EM) algorithm (Dempster et al, 1977). A major advantage of this approach is the use of likelihoods for the probability models and thus access to various likelihood-based model selection criteria to compare different models. Model-based clustering approaches for binary, count and categorical data have been proposed by Biernacki et al (2000), Pledger (2000), Govaert and Nadif (2008), Arnold et al (2010), Labiod and Nadif (2011), Pledger and Arnold (2014). More recently, DeSantis et al (2008), Biernacki and Jacques (2015), Fernández et al (2016), and Matechou et al (2016) have also used these models for ordinal responses in cross-sectional settings.

The purpose of this article is to extend this model-based clustering approach to the case of longitudinal ordinal data. In particular, we propose a latent transitional model that uses the POM parametrisation and includes cluster interactions. Our model contributes to the literature in a number of ways. In contrast to Cheon et al (2014) and Frydman (2005), neither the number of mixture components nor the time transitions are fixed or restricted in any way (other than through identifiability constraints). Importantly, cluster interactions provide a flexible way to capture different time patterns by cluster. On the other hand, unlike Pamminger et al (2010) and Frühwirth-Schnatter et al (2012), our model is time-heterogenous and by using cumulative distributions it is specifically tailored to ordinal data. Finally, our use of the WAIC (Watanabe, 2009) for model comparison of finite mixture models, is also novel as to date in the Bayesian literature it is only being used for mixtures of continuous data (Gelman et al, 2014b; Vehtari et al, 2017).

The structure of this article is as follows: Section 2 describes the data to be used to illustrate the model. Next, Section 3 shows the methodology in detail, including the likelihood function, Bayesian estimation methodology, model comparison, classification strategy and a validation of the model using simulated data.

2 Data

2.1 Self-Reported Health Status over 2001-2011 in Australia

We apply our model to self-reported health status (SRHS) from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. 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. SRHS was collected using the following question: "In general, would you say your health is:" with alternatives: "Poor", "Fair", "Good", "Very Good" and "Excellent". SRHS is thus an ordinal variable with five categories. We use individuals with complete SRHS records over 2001 to 2011. Overall, this amounts to 4,660 respondents over 11 occasions.

Figure 1 presents the distribution of SRHS over the study period. The upper panel shows the SRHS distribution for all years of the study period and the lower panel presents this distribution at the beginning and end. In the upper panel, we can see that there is a general tendency to report slightly lower levels of SRHS over time. With the exception of 2009, every year the SRHS distribution shifts a little to the left, towards the lower end of the ordinal scale. The bottom panel allows us to have a closer look at the beginning and end of the study period. In 2001, most individuals reported "Very Good" health. This was very closely followed by "Good" SRHS. 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 slightly lower health levels and most people reported being in "Good" health. At the same time, the number of "Excellent" and "Very Good" answers decreased and "Poor" and "Fair" increased. In summary, responses were slightly less positive but otherwise similar during 2001-2011.

For each individual SRHS is highly correlated across time. Table 1 presents the 2001-2011 transitions between ordinal categories for all individuals. Proportions in the diagonal and its adjacent cells are very high, about 40% or more. This means that even after 11 years individuals in this survey are very likely to report a health status that is very similar to their starting one. Put simply, SRHS was fairly stable between 2001 and 2011, a fact that confirms what we already observed in Figure 1.

			2011					
		% 2001	Poor	Fair	Good	V. Good	Excellent	Total
2001	Poor	0.02	0.42	0.40	0.14	0.04	0.00	1.00
	Fair	0.13	0.13	0.44	0.34	0.07	0.01	1.00
	Good	0.32	0.02	0.21	0.54	0.20	0.02	1.00
	V. Good	0.37	0.01	0.09	0.38	0.46	0.07	1.00
	Excellent	0.16	0.01	0.04	0.21	0.47	0.27	1.00
	Total	1.00						

Table 1: 2001-2011 transitions in self-reported health status (SRHS) in the HILDA survey.

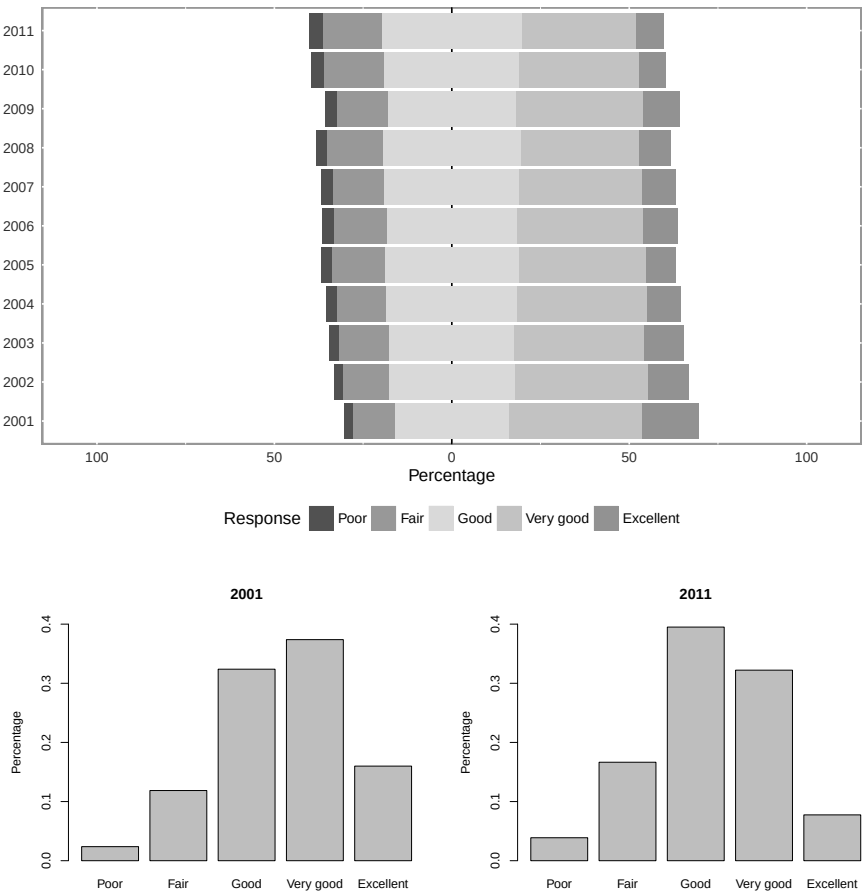

Fig. 1: Distribution of self-reported health status (SRHS) in HILDA over 2001-2011. The upper panel shows stacked bar charts for all years and the lower panels show barplots for 2001 and 2011.

3 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. We further assume the existence of R clusters of individuals but the cluster membership is unknown. Subjects come from latent cluster r with probability $\pi_r \geq 0$, $\sum_{r=1}^{R} \pi_r = 1$ and let $P(Y_{ij} = k | i \in r, y_{i(j-1)} = k') = \theta_{rjk'k}$, where $i \in r$ indicates the membership of subject i is cluster r. We extend the POM by modelling the cumulative probability of each ordinal outcome as

$$\text{Logit}[P(Y_{ij} \leq k | i \in r, y_{i(j-1)})] = \mu_k - \alpha_r - \sum_{k'=1}^{q} \beta_{rk'} I(y_{i(j-1)} = k') - \gamma_j$$

where $I(.)$ is an indicator function equal to 1 if the argument is true. This model could be expressed equivalently as:

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

That is, each ordinal response y_{ij} is the realization of a categorical distribution with probabilities $\theta_{rjk'1}, \ldots \theta_{rjk'q}$. Notice that the linear predictor for the probability $\theta_{rjk'k}$ 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 sometimes referred as a cut point for each ordinal category, α_r is the effect of the latent cluster r, $\beta_{rk'}$ the effect of having an outcome k' at the previous occasion for subjects in cluster r, and γ_j the effect of occasion j. The choice of a negative sign preceding α_r, $\beta_{rk'}$, and γ_j 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).

Importantly, the cluster interactions provide flexible and parsimonious ways to introduce different time patterns by cluster. They allow both time-constant and time-varying unobserved heterogeneity to be captured. The inclusion of γ_j allows $\theta_{rjk'k}$ to vary over time, that is for individuals to have time-heterogeneous transitions between ordinal categories. Note also that this transitional model does not model the first response ($Y_{.1}$) and instead conditions on its value. Finally, notice also that following the seminal paper of Albert and Chib (1995) we set $\mu_1 = 0$ and have no constraint on any α_r. By fixing the first cut point, this parametrisation allows better mixing of the MCMC chain during Bayesian estimation.

3.1 Likelihood

Given the dependence on the previous outcome, we can factorize the likelihood to separate the contribution of the first occasion, e.g. $Y = (Y_{.1}, \widetilde{Y})$. Assuming independence over the rows, the likelihood for the observations with $j \geq 2$ becomes

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

3.2 Bayesian Estimation

Following Robert and Casella (2005),Gelman et al (2014a), McKinley et al (2015) and Fernández and Arnold (2016), we use the following weakly informative priors:

$$\begin{aligned}
\mu \mid \sigma_\mu^2 &\overset{iid}{\sim} \text{OS}[\text{Normal}(0,\sigma_\mu^2)],\ \mu_k > \mu_{k-1}; k = 1,\ldots,q\ \mu_0 = -\infty; \mu_1 = 0, \mu_q = \infty \\
\alpha_r \mid \sigma_\alpha^2 &\overset{iid}{\sim} \text{Normal}(0,\sigma_\alpha^2),\ r = 1,\ldots,R \\
\beta_{rk'} \mid \sigma_{\beta r}^2 &\overset{iid}{\sim} \text{Degenerate Normal}(q;0,\sigma_{\beta r}^2),\ k' = 1,\ldots,q;\ \sum_{k'=1}^{q}\beta_{rk'} = 0; \forall r = 1,\ldots,R \\
\gamma_j \mid \sigma_\gamma^2 &\overset{iid}{\sim} \text{Degenerate Normal}(p-1;0,\sigma_\gamma^2),\ j = 2,\ldots,p,\ \sum_{j=2}^{p}\gamma_j = 0 \\
\sigma_\mu^2 &\sim \text{Inverse Gamma}(a_\mu, b_\mu) \\
\sigma_\alpha^2 &\sim \text{Inverse Gamma}(a_\alpha, b_\alpha) \\
\sigma_{\beta r}^2 &\sim \text{Inverse Gamma}(a_\beta, b_\beta) \\
\sigma_\gamma^2 &\sim \text{Inverse Gamma}(a_\gamma, b_\gamma) \\
\pi &\sim \text{Dirichlet}(\phi),\ r = 1,\ldots,R.
\end{aligned} \tag{3}$$

where OS=Order Statistics and the hyperparameters are set to: $a_\mu = a_\alpha = a_\beta = 3$, $b_\mu = b_\alpha, b_\beta = 40$, and $\phi = 1.5$. In words, we assign Truncated Normal priors for the cut points μ, Normal priors with zero mean and unknown variance for α, Degenerate Normal priors with zero mean and unknown variance for β and γ, Dirichlet prior for the mixing probabilities π, and Inverse Gamma priors for the unknown variances σ_μ^2, σ_α^2 and σ_β^2. Note that, a degenerate normal distribution is a probability distribution with normally distributed realizations whose sum is equal to their mean multiplied by the number of realizations. It is thus a convenient prior for random variables with support in $\mathbb{R}$ and sum to zero constrains such as β and γ in our model. Formal derivation of this prior can be found in Fernández and Arnold (2016). Figure 2 shows a graphical representation of the model and priors.

Posterior distributions for the model parameters are not available in closed form. To perform the posterior computation, we then use a Markov Chain Monte Carlo (MCMC) sampling scheme. In particular, we use a Random-Walk Metropolis-Hastings algorithm (Metropolis et al, 1953; Hastings, 1970) to sample blocks of parameters separately (μ, α, β and π and the parameters of the priors). For instance, to sample from the posterior distribution of $\mu = (\mu_1 = 0, \mu_2 \ldots, \mu_{q-1})$ we follow McKinley

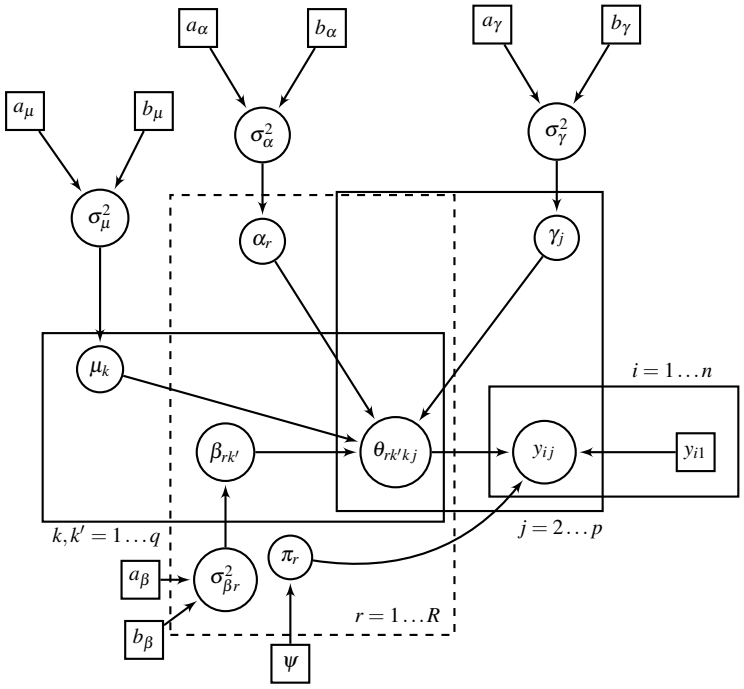

Fig. 2: Graphical representation of the latent transitional model and priors.

et al (2015) and use a truncated uniform with a fixed stepsize τ as a proposal. The algorithm is as follows:

1. Set starting values for all parameters:
   $(\mu, \alpha, \beta, \gamma, \pi, \sigma^2_\mu, \sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma) = (\mu_0, \alpha_0, \beta_0, \gamma_0, \pi_0, \sigma^2_{\mu_0}, \sigma^2_{\alpha_0}, \sigma^2_{\beta_0}, \sigma^2_{\gamma_0})$
2. Set the stepsize of the proposal (τ)
3. Choose k from $k = 2, \ldots, q-1$ at random and generate a new μ'_k candidate from

$$\mu'_k \mid \mu_k, \mu_{k-1}, \mu_{k+1} \sim U[\ \max(\mu_k - \tau, \mu_{k-1}),\ \min(\mu_k + \tau, \mu_{k+1})] \qquad k = 1, \ldots, q-1$$

4. Accept μ'_k with probability

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

5. Repeat steps 3 and 4 until convergence.

Here $P(Y|\mu, \alpha, \beta, \gamma, \pi, Y_{\cdot 1})$ represents the likelihood in equation 2 and $P(\mu|\sigma^2_\mu)$ is the prior for parameters: $\mu \mid \sigma^2_\mu \overset{iid}{\sim} \text{OS}[\text{Normal}(0, \sigma^2_\mu)]\ \mu_k > \mu_{k-1},\ k = 1, \ldots, (q-1)$. Detailed proposals for all model parameters are given in Appendix A.

3.3 Model Comparison

There are several ways to compare models in a Bayesian framework: (i) using Bayes Factors (Kass and Raftery, 1995), (ii) estimating the joint posterior distribution of all competing models using Reversible Jump MCMC (Green, 1995; Richardson and 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-like) 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 specially 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 and Peel, 2000; Richardson and Green, 1997; Marin et al, 2005).

To compare among competing models we therefore use the Widely Aplicable Information Criterion (WAIC) developed by Watanabe (2009) which uses the posterior distribution of all the parameters. For a model with parameters ω and data Y, the WAIC is defined as

$$\begin{aligned} \text{WAIC}_1 &= -2\sum_{i=1}^{n}\log\int p(Y_i|\omega)p(\omega|Y)d(\omega) + 2p_1 \\ &\approx -2\sum_{i=1}^{n}\log\left[\frac{\sum_{s=1}^{S}p(Y_i|\omega^s)}{S}\right] + 2p_1 \end{aligned} \tag{4}$$

where S is the number of iterarions in the MCMC chain and p_1 is the effective number of parameters

$$\begin{aligned} p_1 &= \sum_{i=1}^{n}\{\log\int p(Y_i|\omega)p(\omega|Y)d(\omega) - \int \log p(Y_i|\omega)p(\omega|Y)d(\omega)\} \\ &\approx \sum_{i=1}^{n}\{\log\left[\frac{\sum_{s=1}^{S}p(Y_i|\omega^s)}{S}\right] - \left[\frac{\sum_{s=1}^{S}\log p(Y_i|\omega^s)}{S}\right]\} \end{aligned} \tag{5}$$

Alternatively, the effective number of parameters and the WAIC can also be approximated by

$$\begin{aligned} p_2 &= \sum_{s=1}^{S}\text{Var}[\log p(Y_i|\omega^s)] \\ \text{WAIC}_2 &\approx -2\sum_{i=1}^{n}\log\left[\frac{\sum_{s=1}^{S}p(Y_i|\omega^s)}{S}\right] + 2p_2 \end{aligned} \tag{6}$$

Defined in these ways the WAIC is on the same scale as the AIC and BIC. The term $p(Y_i|\omega)$ is the contribution of the ith observation to the likelihood and is referred to as *pointwise predictive density* in the literature (Geisser and Eddy, 1979; Gelman et al, 2014a). We follow this terminology here and call the first component of the WAIC definition *log predictive density* (LPD).

It is important to stress that the WAIC overcomes the label switching problem by integrating out the posterior distribution of all parameters $p(\omega|Y)d(\omega)$ from the pointwise predictive density $p(Y_i|\omega)$. In practice, this integral is approximated by Monte Carlo integration using all MCMC draws $p(Y_i|\omega^s)$ as shown in the second line of (4). A similar procedure is used to approximate the integrals involved in the calculation of p_1 (5).

As a comparison, we also present the Deviance Information Criterion (DIC) by Spiegelhalter et al (2002, 2014), which is being used extensively in Bayesian applications. We separate out its two components: posterior mean Deviance ($\overline{D}$) and number of effective parameters (p_{DIC}) so that these can be adequately compared to the WAIC components. The DIC is defined as:

$$\begin{aligned} &\text{DIC} = \overline{D(\omega)} + p_{DIC} \\ &\text{where:} \\ &\overline{D(\omega)} = -2\mathbf{E}_{\omega|Y}[\log p(Y|\omega)] = -2\left[\frac{\sum_{s=1}^{S}\log p(Y|\omega^s)}{S}\right] \\ &p_{DIC} = -2\mathbf{E}_{\omega|Y}[\log p(Y|\omega)] + 2\log[p(Y|\tilde{\omega}(Y))] \end{aligned} \tag{7}$$

Note that DIC requires a plug-in estimate of the posterior distribution $\tilde{\omega}(Y)$ to be calculated. Here we take the mean posterior for each parameter $\tilde{\omega}(Y) = \mathbf{E}[\omega|Y]$ as a plug-in but it could also be any value that adequately represent the posterior distribution such as the median or the mode.

DIC must be used with caution in cases where the posterior distribution of the parameters is multimodal, such as in mixtures or hierarchical models. In these cases, the effective number of parameters p_d can be negative and thus the resulting DIC value should not be trusted (Celeux et al, 2006; Spiegelhalter et al, 2014). Moreover, DIC is not asymptotically consistent as it is not aiming to select the *true model* (Spiegelhalter et al, 2002).

3.4 Model validation using simulated data

In order to validate the model, we simulated data from the mixture model in (1). Specifically, we used a mixture with three-components, equal proportions, five ordinal categories and sampled 1000 observations over 15 occasions ($R = 3, n = 1000, p = 15, q = 5$). Varying cluster-occasion interactions ($\beta_{rk'}$) were also setup so that the model exhibits different patterns over time. True values for all parameters ($\mu, \alpha, \beta, \gamma, \pi$) can be found in Table 2. This model is estimated using three parallel MCMC chains, each with a burn-in of 27,000 and length 540,000. We assess the convergence of the MCMC chains using the Potential Scale Reduction Factor (PSRF) by Gelman and

Rubin (1992). PSRF is a multi-chain diagnostic test with values higher than 1.2 indicating lack of convergence. For this syntetic dataset, PSRF values for all parameters are very close one (not shown here) and thus we can conclude that our MCMC chains have converged.

Table 2 shows summary statistics for the posterior distribution of all model parameters. In addition to the true values, it shows the mean posterior, standard error (SE) and the 95% credible interval. Importantly, the mean posterior for all parameters is very close to their true values (given their SE's). Moreover, in all cases the 95% credible intervals contain the true values of the parameters. These results are reassuring because they provide evidence that the proposed model and MCMC sampler are working properly. `R` and `C++` scripts to completely reproduce the simulation results in Table 2 are publicly available at: `https://github.com/Cholokiwi/pomtc`.

3.5 Classification

We use heatmaps to visually assess the fuzziness of the estimated classification of individuals into clusters. It is important to stress that heatmaps should only be used to visualise the best fitting model(s) after comparison among all candidate models as the human eye tends to see patterns in any given image (Wilkinson and Friendly, 2009).

Classification probabilities $\hat{z}_{ir}$ close to one would mean that our fuzzy probabilistic clustering is "crisp". To do so, we calculate the co-clustering probabilities for all individuals. We define here a co-clustering probability $C_{ii'}$ as 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 at the MCMC iteration s:

$$C_{ii'} = \frac{\sum_{s=1}^{S}\sum_{r=1}^{R}\hat{z}_{ir}^{s}\hat{z}_{i'r}^{s}}{S}, \text{ for } i,i' = 1,\ldots,n \text{ and } s = 1,\ldots,S. \tag{8}$$

where

$$\hat{z}_{ir}^{s} = \frac{\pi_r^s \prod_{j=2}^{p}\prod_{k'=1}^{q}\prod_{k=1}^{q}\theta_{rjk'k}^{s,I(y_{ij}=k)}}{\sum_{a=1}^{R}\pi_a^s \prod_{j=2}^{p}\prod_{k'=1}^{q}\prod_{k=1}^{q}\theta_{ajk'k}^{s,I(y_{ij}=k)}}$$

That is, $\hat{z}_{ir}$ is the posterior mean of the classification probabilities z_{ir}^{s} over the MCMC chain. Note that $\theta_{rjk'k}^{s,I(y_{ij}=k)}$ is obtained evaluating (1) at the model parameters $\mu,\alpha,\beta,\gamma,\pi$ for each MCMC iteration s.

4 Results

We illustrate the model using a random subsample of 230 individuals from HILDA who had complete responses over 2001-2011, that is individuals with SRHS in all eleven waves. We used the `R` statistical language, version `3.3.3` (R Core Team, 2017), linked with `C++` routines to implement the model. The model is fitted using

				95% Credible Interval	
Par	True	Mean	SE	Lower	Upper
μ_2	0.98	1.04	0.03	0.98	1.08
μ_3	1.79	1.84	0.03	1.78	1.90
μ_4	2.77	2.82	0.04	2.74	2.90
α_1	1.39	1.42	0.06	1.30	1.53
α_2	0.39	0.37	0.05	0.27	0.48
α_3	2.39	2.33	0.09	2.15	2.50
β_{11}	-0.53	-0.57	0.07	-0.72	-0.43
β_{12}	-0.46	-0.39	0.07	-0.53	-0.26
β_{13}	-0.30	-0.31	0.07	-0.45	-0.17
β_{14}	-0.30	-0.33	0.07	-0.47	-0.18
β_{15}	1.59	1.61	0.08	1.45	1.76
β_{21}	-0.98	-0.96	0.06	-1.07	-0.84
β_{22}	-0.89	-0.90	0.08	-1.06	-0.75
β_{23}	-0.88	-1.06	0.10	-1.24	-0.85
β_{24}	-0.56	-0.53	0.11	-0.73	-0.31
β_{25}	3.30	3.45	0.12	3.20	3.65
β_{31}	-1.36	-1.45	0.14	-1.74	-1.17
β_{32}	-0.30	-0.42	0.13	-0.67	-0.17
β_{33}	-0.29	-0.29	0.11	-0.52	-0.09
β_{34}	-0.02	0.05	0.10	-0.14	0.23
β_{35}	1.96	2.10	0.08	1.93	2.25
γ_2	0.21	0.20	0.06	0.08	0.31
γ_3	-0.18	-0.17	0.06	-0.29	-0.05
γ_4	0.58	0.63	0.06	0.51	0.76
γ_5	-0.35	-0.33	0.06	-0.45	-0.21
γ_6	0.15	0.20	0.06	0.09	0.32
γ_7	0.68	0.61	0.06	0.49	0.73
γ_8	0.77	0.75	0.06	0.63	0.88
γ_9	-1.11	-1.10	0.06	-1.21	-0.97
γ_{10}	-0.35	-0.42	0.06	-0.53	-0.29
γ_{11}	-1.22	-1.22	0.07	-1.35	-1.10
γ_{12}	-0.17	-0.13	0.06	-0.24	-0.01
γ_{13}	1.56	1.60	0.07	1.48	1.73
γ_{14}	1.72	1.68	0.07	1.54	1.84
γ_{15}	-2.28	-2.30	0.07	-2.44	-2.17
π_1	0.33	0.35	0.03	0.29	0.41
π_2	0.33	0.35	0.02	0.31	0.38
π_3	0.33	0.30	0.03	0.25	0.36
log-like	-16806	-16809	4.21	-16817	-16801

Table 2: True values and posterior summary statistics for model parameters in simulated data ($n = 1000, p = 15, q = 5, R = 3$).

a varying number of latent groups from one ($R = 1$ no-clustering) to seven ($R = 7$ latent groups). For each R, three parallel chains with different starting points were run for 4.5 million iterations, thin by 500 and the first three quarters were discarded as burn-in. Thus, inference was carried out using $S = 6750$ iterations ($3 \times 4.5 \times 10^6 \times 0.25/500 = 6750$). Depending on the number of mixture components, each chain took around 15-60 minutes to run using Xeon E5-2680 2.50GHz CPUs. After selecting the best fitting model using the WAIC, we post-processed the chains according to the relabelling algorithm of Stephens (2000) to deal with label switching. Finally, to ease comparability we also sorted the clusters by increasing cluster effect α so that respondents in cluster 1 have the lowest levels of SRHS and those in cluster R the highest ones.

Table 3 shows the model comparison results. For each fitted model, it presents: number of clusters (R), number of parameters (npars), mean posterior of the log-

likelihood ($\overline{\log l}$), DIC ($\overline{D}$ and p_d) and the two versions of the WAIC and their corresponding components (LPD, p_1, p_2, WAIC_1, and WAIC_2). The table allows us to highlight a few things. Firstly, the model with six clusters have the lowest values for both WAIC versions and thus provide the best fit. Secondly, although $\overline{\log l}$, $\overline{D}$ and WAIC decrease monotonically until $R = 6$, these decrements are very are small from $R = 3$ onwards. In other words, a model with $R = 6$ provides only a slightly better fit than a model with $R = 5$ and so on. Thus, an information criterion with a higher penalty for the number of parameters will be likely to choose a model with fewer mixture components. Last but not least, DIC has a negative number of effective parameters p_d and thus should not be used for model comparison in this case. Neither version of the WAIC has this drawback.

Posterior summary statistics, convergence diagnostics, traceplots and posterior distributions for the above model with $R = 6$ can be found in Appendices C and D.

R	npars	$\overline{\log l}$	$\overline{D}$	p_{DIC}	DIC	LPD	p_1	WAIC_1	p_2	WAIC_2
1	20	-2354	4708	17	4724	4682	25	4733	26	4734
2	27	-2195	4390	23	4412	4360	29	4419	30	4421
3	33	-2136	4272	-466	4297	4238	34	4306	35	4308
4	39	-2129	4258	-51	4286	4222	36	4293	37	4296
5	45	-2124	4248	-412	4276	4206	42	4289	44	4294
6	51	-2121	4242	-80	4271	4202	40	4282	42	4286
7	57	-2122	4244	-2428	4250	4200	44	4288	47	4294

Table 3: Bayesian model comparison using WAIC for the HILDA dataset.

Next, we check the classification results using the co-clustering probabilities, i.e. probability that respondents belong to the same cluster over all MCMC iterations, defined in (8) for the model with six components. Figure 3 displays these co-clustering probabilities for the model in the original data (top panel) and ordered by cluster (bottom panel). We can see that when ordering the respondents by cluster, we are able to visualise high co-clustering probabilities within cluster as well as their relative size, area of rectangles in the diagonal, which provides a visual indication of the estimated cluster proportions $\hat{\pi} = (0.08, 0.32, 0.26, 0.24, 0.05, 0.04)$. The selected model with six clusters not only provides the best fit among $R = 1, \ldots, 7$ but also provides a crisp allocation of individuals to the estimated clusters.

What do these estimated clusters look like? Figure 4 displays the overall and cluster specific distribution of SRHS by year. To ease interpretation, the plots include point estimates for selected parameters. Values for $\hat{\alpha}_r$ and $\hat{\pi}_r$ correspond to posterior means and $\hat{\sigma}^2_{\beta r}$ to posterior medians. Furthermore, we also sorted the latent groups by increasing cluster effect $\hat{\alpha}_r$ so that individuals in the first and last cluster have the lowest and highest levels of SRHS, $\hat{\alpha}_1 = 2.8$ and $\hat{\alpha}_6 = 12.4$ respectively. Plots also include an estimate of the variance of the cluster interactions, posterior median $\hat{\sigma}^2_{\beta r}$ which measures the mobility between ordinal categories for individuals that belong to the cluster. Higher values of $\hat{\sigma}^2_{\beta r}$ imply a cluster formed by individuals that move more between ordinal categories over time. That is, respondents in these clusters tend

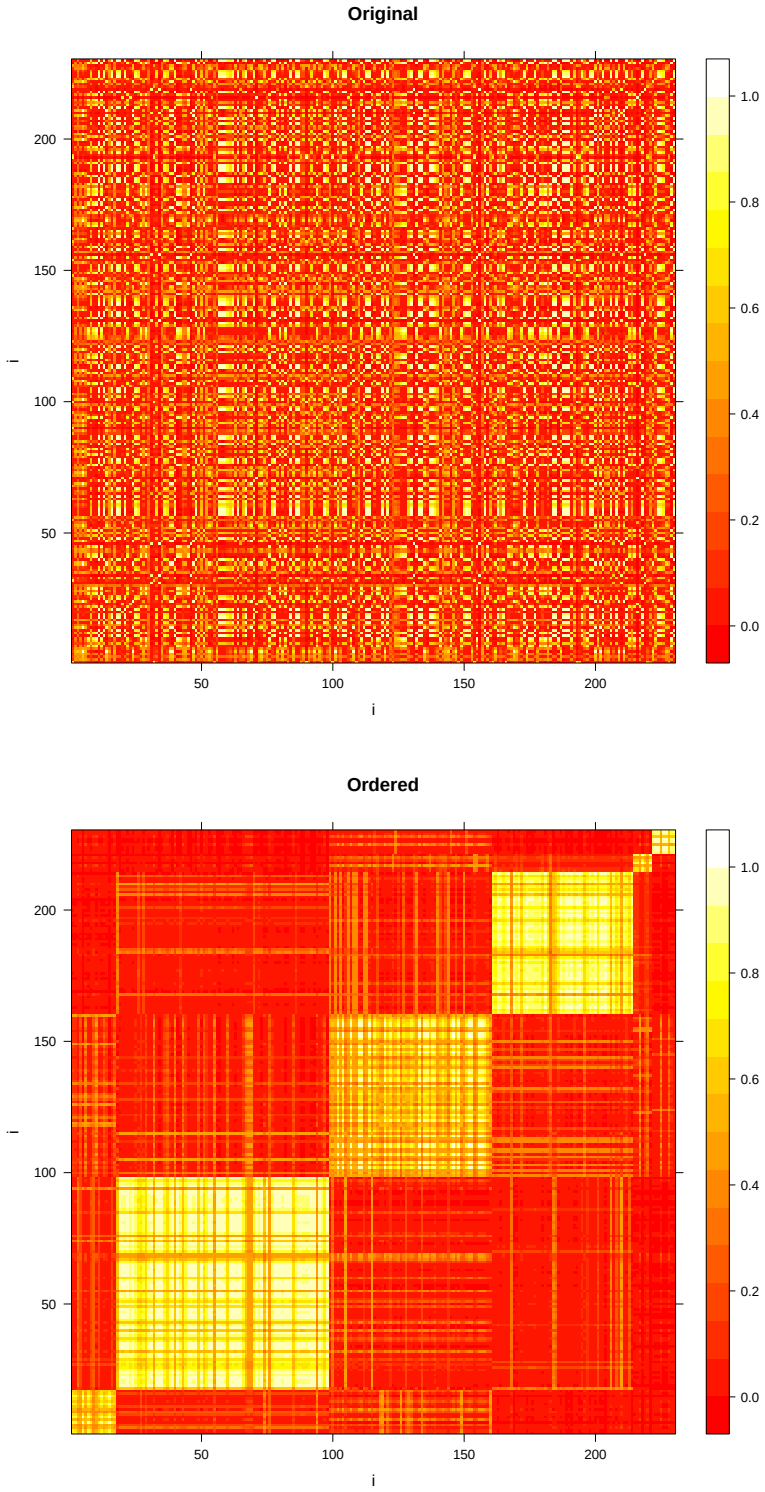

Fig. 3: Co-clustering probabilities for the original HILDA data (upper panel) and the model with six latent groups (bottom panel).

to change more their health status from 2001 to 2011. Conversely, smaller $\hat{\sigma}^2_{\beta r}$ values imply a cluster where individuals have less movement among ordinal levels and thus tend to report a similar health status in the study period.

Figure 4 shows very different SRHS cluster profiles over time. Firstly, cluster 3 has the highest mobility, $\hat{\sigma}^2_{\beta 3} = 8.3$, and spreads among all five ordinal categories. Respondents in this cluster are about a quarter of the total ($\hat{\pi}$=0.26). Although less pronounced, Cluster 2 exhibits a high pattern of mobility $\hat{\sigma}^2_{\beta 2} = 0.8$ and accounts for a third of the sample ($\hat{\pi}$=0.32). Despite being located at different levels of the ordinal scale ($\hat{\alpha}$), the remaining clusters have lower levels of mobility with $\hat{\sigma}^2_{\beta r} \approx 0.3$. People in cluster 1 for instance have a neutral perception of their health status (centered around the "Fair" category). On the other hand, cluster 6 represents the extreme of positiveness as respondents in these are extremely satisfied with their health status and have responses mostly in the "Excellent" category.

5 Discussion and Conclusions

Model-based clustering approaches provide a way to identify latent groups and reduce the data dimensionality. In this paper, we proposed a latent transitional model that uses the proportional odds parametrisation for longitudinal ordinal data. The proposed finite mixture model includes cluster (α_r) and occasion (γ_j) effects as well as cluster interactions with the immediate past response ($\beta_{rk'}$) which allow time-heterogeneous transitions and the lagged response to have a different effect on each cluster. We estimated the model within a Bayesian approach using MCMC with a block Metropolis-Hastings sampler. To compare among models with different number of mixture components we used WAIC but also shown DIC for completeness. In addition to that, a relabelling strategy allowed us to identify the latent groups itself. We applied the model to self-reported health status data ("Poor", "Fair", "Good", "Very Good" and "Excellent") over 11 years in an Australian household panel survey and found evidence for six latent groups with distinct patterns over time. Our proposal extends the currently available models by providing a flexible way to capture different time patterns by cluster and time-heterogeneous transitions. By using the WAIC to compare the number of components in finite mixtures of ordinal data, it also contributes to the Bayesian literature of model comparison.

Our model has several limitations. Firstly, it is computer-intensive and estimation might become impractical when dealing with big datasets, e.g. millions of units followed over hundreds of occasions. 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 it using the Metropolis-Hastings sampler. This however is only a technological limitation and can be overcome, or at least alleviated, by the use of grid computing and parallelizing the computer code used for estimation. Alternative ways to deal with big data in our context will be to resort to non-exact Bayesian methods such as Variational Approximations (Wainwright and Jordan, 2008; Hui et al, 2017) and Approximate Bayesian Computation (Beaumont et al, 2002; Gutmann et al, 2018).

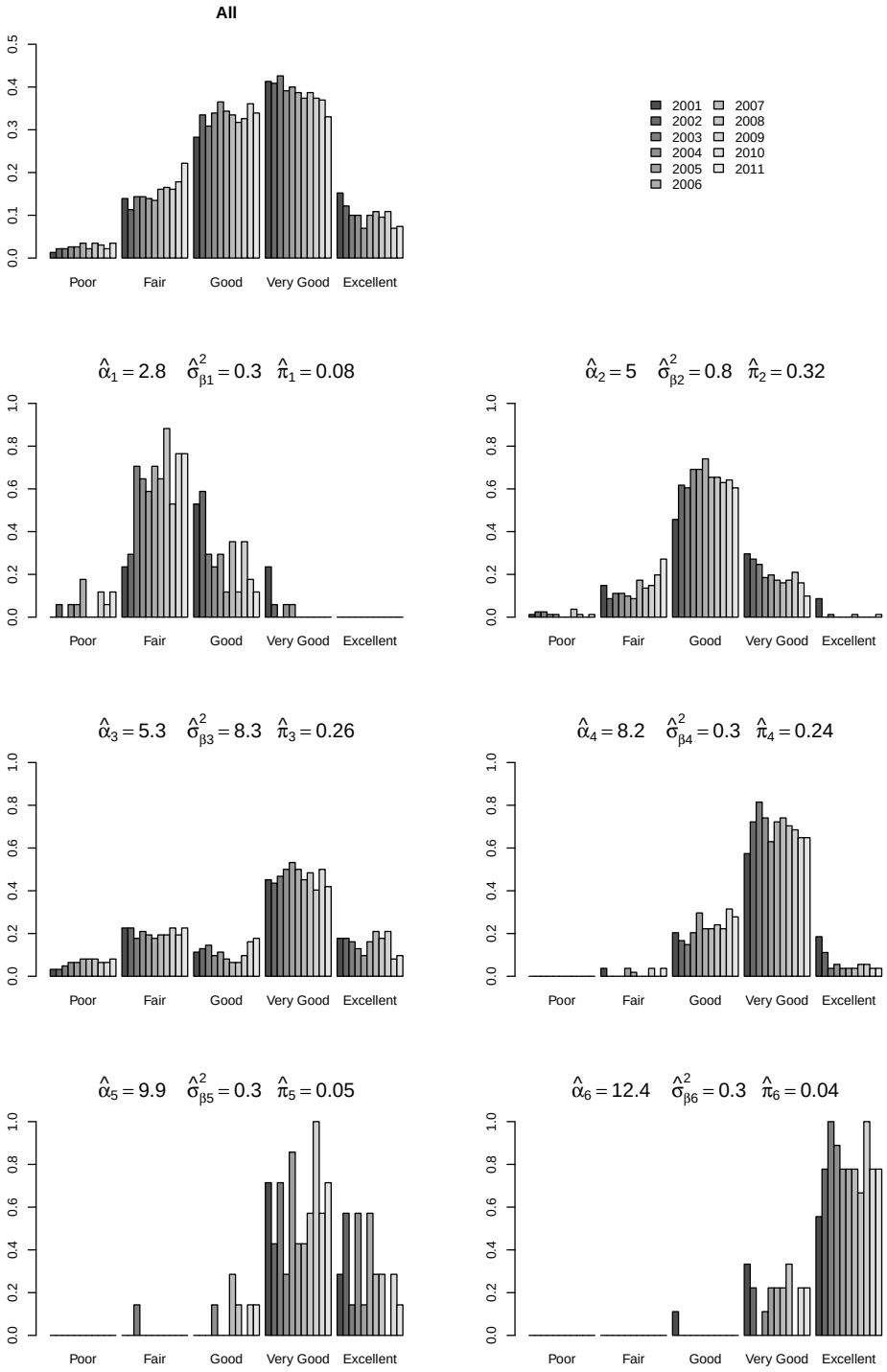

Fig. 4: Distribution of SRHS by year in all HILDA data and estimated six clusters. Values for $\hat{\alpha}_r$ and $\hat{\pi}_r$ correspond to posterior means and $\hat{\sigma}^2_{\beta r}$ to posterior medians.

Secondly, caution should be taken when interpreting the estimated mixture components. Mixture models are very flexible and can fit any dataset given enough components. Therefore, selecting too many cluster components is a danger when working with mixture models. Information criteria like WAIC and DIC penalise model complexity but as the Bayesian equivalents of the AIC they also have a "conservative" penalty term (twice the number of effective parameters) which does not take into account sample size. In fact, this might be the case for some of the estimated clusters in the health status application presented here. For instance, clusters 5 and 6 (Figure 4) have cluster interactions with similar variances but centered on more positive levels of the ordinal scale. Furthermore, models with $R = 3$ and $R = 6$ have very similar mean posterior log-likelihoods, -2136 and -2121 in Table 3, and thus provide similar estimated parameters and clusters. Therefore, other non-predictive Bayesian information criteria like the WBIC (Watanabe, 2013; Friel et al, 2017) or the sBIC (Drton and Plummer, 2017) could be worth exploring here.

Lastly, Bayesian approaches can always be sensitive to choice of priors. In order to check for this possibility, in Section 3.4 we use simulated data to validate the model and found that our weakly informative priors and MCMC sampler were able to recover the true parameters of the model. It would be interesting though, to run a simulation study where our Bayesian estimation strategy is further tested in a variety of scenarios. Nonetheless, no simulation study will provide an absolute guarantee that either the proposed priors or the model itself are appropiatte for every data application.

We plan to extend the model in two directions: exploring other ways to incorporate the correlation and including the number of mixture components as parameters in the model. The former could be done by including past responses of higher orders, not just the previous response as in the current model. The latter would imply the use of an encompassing model, where the number of mixture components is not longer fixed. Reversible Jump MCMC (Green, 1995; Richardson and Green, 1997) and Bayesian Non-Parametric models (Müller et al, 2015; DeYoreo and Kottas, 2017) are for instance examples of such trans-dimensional approaches. Albeit having a more complex parameter space, these models also estimate the distribution for the number of mixture components and thus simplify the comparison amongst competing models.

Acknowledgements The work is being supported by the Marsden Fund Grant 16-VUW-062 from the Royal Society of New Zealand. We also would like to thank Shirley Pledger from VUW for many useful discussions.

This paper uses unit record data 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.

More information about the HILDA survey can be found at:
`https://www.melbourneinstitute.com/hilda/`

References

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

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

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

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

Beaumont MA, Zhang W, Balding DJ (2002) Approximate bayesian computation in population genetics. Genetics 162(4):2025–2035

Biernacki C, Jacques J (2015) Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Statistics and Computing pp 1–15

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

Celeux G, Forbes F, Robert CP, Titterington DM, et al (2006) Deviance information criteria for missing data models. Bayesian analysis 1(4):651–673

Cheon K, Thoma ME, Kong X, Albert PS (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

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal statistical Society 39(1):1–38

DeSantis SM, Houseman EA, Coull BA, Stemmer-Rachamimov A, Betensky RA (2008) A penalized latent class model for ordinal data. Biostatistics 9(2):249–262

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

Diggle PJ, Heagerty PJ, Liang KY, Zeger SL (2002) Analysis of Longitudinal Data, 2nd Edition. Oxford University Press

Drton M, Plummer M (2017) A bayesian information criterion for singular models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79(2):323–380

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

Fernández D, Arnold R (2016) Mode selection for mixture-based clustering for ordinal data. Australian and New Zealand Journal of Statistics 58(4):437–472

Fernández D, Arnold R, Pledger S (2016) Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis 93:46–75

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association 97(458):611–631

Friel N, McKeone J, Oates CJ, Pettitt AN (2017) Investigation of the widely applicable bayesian information criterion. Statistics and Computing 27(3):833–844

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 WF (1979) A predictive approach to model selection. Journal of the American Statistical Association 74(365):153–160

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Statistical Science 7(4):457–472

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014a) Bayesian Data Analysis, 3rd edition. Taylor & Francis

Gelman A, Hwang J, Vehtari A (2014b) Understanding predictive information criteria for bayesian models. Statistics and Computing 24(6):997–1016

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

Green PJ (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82(4):711–732

Gutmann MU, Dutta R, Kaski S, Corander J (2018) Likelihood-free inference via classification. Statistics and Computing 28(2):411–425

Hastings WK (1970) Monte carlo sampling methods using markov chains and their applications. Biometrika 57(1):97–109

Hui FKC, Warton DI, Ormerod JT, Haapaniemi V, Taskinen S (2017) Variational approximations for generalized linear latent variable models. Journal of Computational and Graphical Statistics 26(1):35–43

Kass RE, Raftery AE (1995) Bayes factors. Journal of the american statistical association 90(430):773–795

Kaufman L, Rousseeuw PJ (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

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

Liu I, Agresti A (2005) The analysis of ordered categorical data: An overview and a survey of recent developments. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research 14(1):1–73

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Cam LML, Neyman J (eds) Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, University of California Press, pp 281–297

Manly BF (2005) Multivariate statistical methods: a primer. CRC Press

Marin JM, Mengersen K, Robert CP (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 To appear

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

McCullagh P, Nelder JA (1989) Generalized Linear Models, 2nd edn. London: Chapman & Hall

McKinley TJ, Morters M, Wood JL, et al (2015) Bayesian model choice in cumulative link ordinal regression models. Bayesian Analysis 10(1):1–30

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

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. The journal of chemical physics 21(6):1087–1092

Müller P, Quintana F, Jara A, Hanson T (2015) Bayesian nonparametric data analysis. Springer Series in Statistics

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

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

R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`

Richardson S, Green PJ (1997) On bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society Series B (Methodological) pp 731–792

Robert CP, Casella G (2005) Monte Carlo Statistical Methods (Springer Texts in Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA

Spiegelhalter DJ, Best NG, Carlin BP, 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

Spiegelhalter DJ, Best NG, Carlin BP, Linde A (2014) The deviance information criterion: 12 years on. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(3):485–493

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

Stevens S (1946) On the theory of scales of measurement. Science 103(2684):677–680

Vehtari A, Gelman A, Gabry J (2017) Practical bayesian model evaluation using leave-one-out cross-validation and waic. Statistics and Computing 27(5):1413–1432

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

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

Watanabe S (2013) A widely applicable bayesian information criterion. The Journal of Machine Learning Research 14(1):867–897

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

Appendices

A. Proposals

After choosing initial values for all model parameters (μ, α, β, γ, π, σ^2_μ, σ^2_α, σ^2_β, and σ^2_γ), we proceed to update them according to the following:

$$\mu'_k \mid \mu_k, \mu_{k-1}, \mu_{k+1} \sim U[\ \max(\mu_k - \tau, \mu_{k-1}),\ \min(\mu_k + \tau, \mu_{k+1})]\ k = 2, \ldots, q-1,\ \mu_0 = -\infty, \mu_1 = 0\, \mu_q = \infty$$

$$\alpha'_r \mid \alpha_r \overset{iid}{\sim} \text{Normal}(\alpha_r, \sigma^2_{\alpha p}) \qquad r = 1 \ldots R,$$

$$\beta'_{rk'} \mid \beta_{rk'} \overset{iid}{\sim} \text{Normal}(\beta_{rk'}, \sigma^2_{\beta p}) \qquad k' = 1 \ldots q-1,\ \beta_{rq} = -\sum_{k'=1}^{q-1} \beta_{rk'},\ \forall r = 1 \ldots R,$$

$$\gamma'_j \mid \gamma_j \overset{iid}{\sim} \text{Normal}(\gamma_j, \sigma^2_{\gamma p}) \qquad j = 2 \ldots p-1,\ \gamma_p = -\sum_{j=2}^{p-1} \gamma_j$$

$$\text{logit}(w') \mid \text{logit}(w) \sim \text{Normal}(\text{logit}(w), \sigma^2_{\pi p}) \qquad w = \pi_{r1}/(\pi_{r1} + \pi_{r2})\ r1, r2 \in 1 \ldots R$$

$$\pi'_{r1} = w'(\pi_{r1} + \pi_{r2}) \qquad \pi'_{r2} = (1 - w')(\pi_{r1} + \pi_{r2})$$

$$\log(\sigma'^2_\mu) \mid \log(\sigma^2_\mu) \sim \text{Normal}(\log(\sigma^2_\mu), \sigma^2_{\sigma\mu p})$$

$$\log(\sigma'^2_\alpha) \mid \log(\sigma^2_\alpha) \sim \text{Normal}(\log(\sigma^2_\alpha), \sigma^2_{\sigma\alpha p})$$

$$\log(\sigma'^2_\beta) \mid \log(\sigma^2_\beta) \sim \text{Normal}(\log(\sigma^2_\beta), \sigma^2_{\sigma\beta p})$$

$$\log(\sigma'^2_\gamma) \mid \log(\sigma^2_\gamma) \sim \text{Normal}(\log(\sigma^2_\gamma), \sigma^2_{\sigma\gamma p})$$

With proposals "steps": $\tau = 0.5$, $\sigma^2_{\alpha p} = 0.1$, $\sigma^2_{\beta p} = 0.1$,$\sigma^2_{\gamma p} = 0.1$, $\sigma^2_{\pi p} = 0.25$, $\sigma^2_{\sigma\mu p} = \log(2)$, $\sigma^2_{\sigma\alpha p} = \log(4)$, $\sigma^2_{\sigma\beta p} = \log(1.5)$ and $\sigma^2_{\sigma\gamma p} = \log(2)$

B. Posterior summary statistics and convergence diagnostics, HILDA $R = 6$.

Par	Median	Mean	SE	Lower CI	Upper CI	PSRF
μ_2	3.53	3.54	0.24	3.12	4.04	1.00
μ_3	6.86	6.87	0.27	6.35	7.38	1.00
μ_4	11.06	11.06	0.33	10.46	11.73	1.00
σ^2_μ	0.27	0.34	0.25	0.08	0.74	1.00
α_1	2.76	2.76	0.49	1.74	3.67	1.00
α_2	5.02	4.99	0.33	4.40	5.60	1.04
α_3	5.27	5.27	0.24	4.77	5.74	1.00
α_4	8.20	8.17	0.53	7.47	9.15	1.10
α_5	9.87	9.84	0.73	8.39	11.23	1.07
α_6	12.39	12.40	1.02	10.17	14.64	1.06
σ^2_α	28.60	31.96	15.05	10.72	59.19	1.00
β_{11}	-0.15	-0.17	0.42	-0.97	0.68	1.03
β_{12}	-0.31	-0.31	0.36	-1.01	0.37	1.00
β_{13}	0.23	0.24	0.34	-0.40	0.93	1.01
β_{14}	0.17	0.19	0.44	-0.72	1.06	1.00
β_{15}	0.06	0.05	0.91	-1.86	1.78	1.02
β_{21}	-0.53	-1.92	2.74	-7.06	0.58	1.00
β_{22}	-1.03	-1.50	1.06	-3.68	-0.35	1.00
β_{23}	0.32	0.17	0.51	-1.05	0.89	1.00
β_{24}	1.44	2.16	1.45	0.90	5.06	1.00
β_{25}	-0.26	1.09	2.72	-1.45	6.08	1.00
β_{31}	-6.06	-4.63	2.78	-7.12	0.50	1.00
β_{32}	-3.08	-2.67	1.04	-3.92	-0.69	1.00
β_{33}	-0.45	-0.42	0.61	-1.53	0.66	1.02
β_{34}	4.36	3.63	1.53	0.85	5.19	1.00
β_{35}	5.51	4.09	2.64	-0.85	6.34	1.00
β_{41}	-0.19	-0.29	0.81	-1.19	0.73	1.18
β_{42}	-0.19	-0.22	0.50	-0.98	0.58	1.13
β_{43}	-0.29	-0.30	0.29	-0.85	0.27	1.03
β_{44}	-0.10	-0.03	0.60	-0.62	0.41	1.25
β_{45}	0.78	0.84	0.78	-0.27	1.88	1.17
β_{51}	0.17	0.17	0.44	-0.70	1.00	1.01
β_{52}	0.29	0.31	0.46	-0.68	1.22	1.03
β_{53}	0.26	0.27	0.40	-0.55	1.04	1.04
β_{54}	0.23	0.24	0.36	-0.46	0.94	1.00
β_{55}	-1.03	-0.99	0.78	-2.44	0.70	1.10
β_{61}	-0.06	-0.05	0.44	-0.94	0.87	1.00
β_{62}	-0.06	-0.06	0.46	-1.00	0.84	1.01
β_{63}	-0.15	-0.16	0.45	-1.07	0.66	1.04
β_{64}	0.04	0.06	0.39	-0.77	0.86	1.00
β_{65}	0.24	0.22	0.69	-1.17	1.57	1.04
$\sigma^2_{\beta 1}$	0.28	0.33	0.20	0.09	0.73	1.02
$\sigma^2_{\beta 2}$	0.77	3.54	5.67	0.12	15.43	1.00
$\sigma^2_{\beta 3}$	8.29	8.56	6.86	0.17	20.43	1.00
$\sigma^2_{\beta 4}$	0.25	0.45	1.51	0.09	0.67	1.31
$\sigma^2_{\beta 5}$	0.29	0.35	0.23	0.09	0.77	1.00
$\sigma^2_{\beta 6}$	0.27	0.33	0.24	0.08	0.72	1.00
γ_2	0.42	0.42	0.14	0.15	0.69	1.00
γ_3	0.17	0.17	0.13	-0.09	0.42	1.00
γ_4	0.04	0.04	0.13	-0.20	0.28	1.00
γ_5	-0.08	-0.08	0.13	-0.33	0.17	1.00
γ_6	0.06	0.06	0.13	-0.19	0.29	1.00
γ_7	0.06	0.06	0.13	-0.18	0.32	1.00
γ_8	-0.02	-0.02	0.12	-0.27	0.20	1.00
γ_9	0.06	0.06	0.13	-0.19	0.31	1.00
γ_{10}	-0.24	-0.24	0.13	-0.48	0.01	1.00
γ_{11}	-0.46	-0.47	0.13	-0.71	-0.21	1.00
σ^2_γ	0.16	0.17	0.07	0.07	0.31	1.00
π_1	0.08	0.08	0.02	0.04	0.13	1.01
π_2	0.32	0.31	0.06	0.20	0.40	1.03
π_3	0.26	0.27	0.05	0.19	0.37	1.00
π_4	0.24	0.24	0.04	0.16	0.33	1.15
π_5	0.05	0.06	0.04	0.02	0.13	1.32
π_6	0.04	0.04	0.01	0.01	0.07	1.08
log-like	-2121	-2121	4.54	-2130	-2113	1.03

C. Traceplots and marginal posterior distributions, HILDA $R = 6$.