A Bayesian Model-Based Approach to Estimate Clusters in Repeated Ordinal Data

Roy Costilla* Ivy Liu* Richard Arnold*

Abstract

Traditional cluster analysis methods used in ordinal data, e.g. k-means, are mostly heuristic and lack statistical inference tools to compare among competing models. To address this, we have developed cluster models based on finite mixtures and applied them for the first time to the case of repeated ordinal data. We estimate them within a Bayesian setting using Markov chain Monte Carlo scheme and Metropolis-Hastings sampling. In particular, we present a hierarchical model with data at 3 levels: clusters, individuals and occasions; where only the latter two are observed.

We illustrate the model using 2001-2011 self-reported health status (SRHS) from the Household, Income and Labour Dynamics in Australia (HILDA). SRHS is an ordinal variable with categories: poor, fair, good, very good and excellent. Overall, we found evidence for five latent groups: two where SRHS remains stable, two where it improves overtime and one where it worsens.The data along with these estimated groups are visualized using heatmaps.

Key Words: Ordinal data, repeated measurements, cluster analysis, Bayesian hierarchical models, WAIC, health status.

1. Introduction

A variable with an ordered categorical scale is called *ordinal* (Agresti, 2010). That is, ordinal data is categorical data where the 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), disease severity (not infected, initial, medium, advanced), agreement with a given statement (strongly disagree, disagree, neutral, agree, strongly agree) and in general any other variable that use the Likert scale.

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, traditional cluster approaches such as hierarchical clustering (Kaufman & Rousseeuw, 1990) , association analysis (Manly, 2005), and partition optimization methods like k-means clustering (Lewis et al., 2003); are not based on likelihoods and thus statistical inference tools are not available and 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.

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 gives the model its name and implies that the odds ratios for describing effects of explanatory variables on the ordinal response

*School of Mathematics and Statistics, Victoria University of Wellington, New Zealand. Corresponding author: roy.costilla@msor.vuw.ac.nz. The work presented here is being supported by a Marsden Grant from the Royal Society of New Zealand. We also would like to thank Shirley Pledger and Daniel Fernandez from VUW for many useful discussions.

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

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. Agresti (2010); Vermunt and Hagenaars (2004) discussed three main approaches to analyse such data: marginal models, subject-specific models and transitional models. Here we develop a model based on the last approach only. Tansitional models include 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 at time t. For example, Kedem and Fokianos (2002) used a cumulative logit transitional model in the context of a longitudinal medical study.

Model-based clustering methods using finite mixtures have been proposed by several authors (McLachlan & Peel, 2000; Everitt, Landau, & Leese, 2001), see literature reviews by Fraley and Raftery (2002); Marin, Mengersen, and Robert (2005); Melnykov and Maitra (2010). Models are often fitted using the Expected-Maximisation algorithm (EM) (Dempster, Laird, & Rubin, 1977) and focus on either continuous, discrete or nominal responses. A major advantage of this approach is the availability of likelihoods, for the probability models, and therefore access to various model selection criteria to evaluate and compare different models. Model-based cluster models for binary, count and categorical data have been proposed by Biernacki, Celeux, and Govaert (2000); Pledger (2000); Govaert and Nadif (2008); Arnold, Hayakawa, and Yip (2010); Labiod and Nadif (2011); Pledger and Arnold (2014). More recently, Fernández, Arnold, and Pledger (2014), and Biernacki and Jacques (2015) have also modelled ordinal responses. Our purpose is to extent these models to the case of repeated ordinal data and estimate them withing a Bayesian approach.

The structure of this document is as follows. Section 2 details data to be used to ilustrate the model. Next, section 3 shows the methodology in detail, including the likelihoods, Bayesian estimation and model comparison. Section 4 presents the results and the estimated transition matrices. Finally, discussions and conclusions are presented in section 5.

2. Data

2.1 Health Status over 2001-2011 in Australia

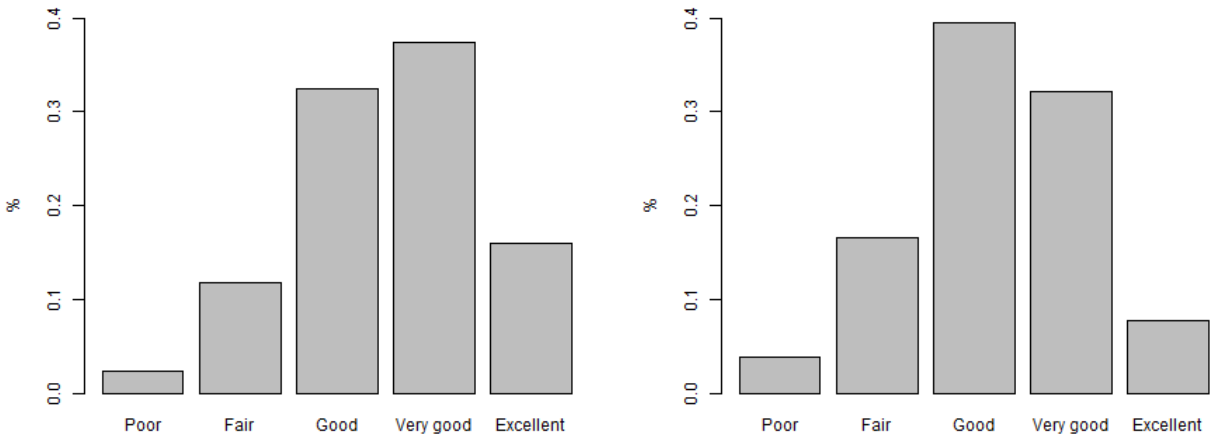
To motivate the model we use self-reported health status (SRHS) from the Household, Income and Labour Dynamics in Australia (HILDA) Survey [1]. HILDA is a household-based panel study which began in 2001 that collects information about economic and subjective well-being, labour market dynamics and family dynamics. The wave 1 panel consisted of 7,682 households and 19,914 individuals.

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

Figure 1 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'

[1]The work presented here 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.

and about a tenth as 'Fair'. A very low number of individuals said their health was 'Poor'. In contrast to that, in 2011 the same individuals reported lower health levels. 'Excellent' and 'Very Good' answers decreased and 'Poor' and 'Fair' increased. The distribution of SRHS shifted to the left and there are fewer responses in the extremes and more in the middle 2011 than in 2001.



2001 2011

Figure 1: Self-Reported Health Status (SRHS) in 2001 and 2011 in HILDA

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 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: 2001-2011 transitions in SRHS

		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

3. Model

Let Y be an ordinal outcome with q levels measured over n subjects on p occasions. Subjects come from one of R latent clusters. The indexes i, r, j are used for subjects, clusters, and occasions; and ordinal levels are denoted k. Extending the POM (McCullagh, 1980),

we model the cumulative probabilities of each ordinal outcome as

$$\text{Logit}[P(Y_{ij} \leq k | i \in r, Y_{i(j-1)})] = \mu_k - \alpha_r - \sum_{k'=1}^{q} \beta_{k'} I(Y_{i(j-1)} = k')$$
$$i = 1 \ldots n; r = 1, \ldots, R; j = 2, \ldots, p; k, k' = 1 \ldots q \quad (1)$$

Where:

- $I(.)$ is an indicator function equal to 1 if the argument is true, and 0 otherwise. Here $Y_{i(j-1)} = k'$ is the value of the response category of the previous ocassion.
- $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{q-1} \leq \mu_q = \infty; k = 1, \ldots, q$
- $\alpha_1 = 0$; $r = 1, \ldots, R$ and
- $\beta_q = 0; k' = 1 \ldots q$

Note that we do not model the first response (Y_{i1}) and instead condition on its value. The μ_k are the cutoff points (POM's latent variable representation), α_r the cluster effects and β_k the effect of having outcome k at the previous ocassion. Finally, π_r represent the mixing probabilities of the finite mixture model ($\pi_r > 0$; $\sum_{r=1}^{R} \pi_r = 1$).

3.1 Likelihood

Given the dependence on the previous outcome, we can factorize the likelihood to separate the contribution of the first occasion ($Y = (Y_{i1}, \widetilde{Y})$). Let π_r be the proportion of subjects from cluster r and $P(Y_{ij} = k | i \in r, Y_{i(j-1)=k'}) = \theta_{rk'k}$. The model's likelihood for the transitions ($j \geq 2$) then becomes

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

3.2 Bayesian Estimation

Following Robert and Casella (2005); Arnold et al. (2010); Gelman et al. (2014); McKinley, Morters, Wood, et al. (2015) we use the following weakly informative priors:

$$y_{ij} \mid \theta_{rk'.}, i \in r \sim \text{Discrete}_q(\theta_{rk'.}),$$
$$i = 1 \ldots n;\ j = 2, \ldots, p;\ r = 1, \ldots, R;\ k' = 1 \ldots q$$
$$\theta_{rk'k} \mid \mu, \alpha, \beta = \frac{1}{1 + e^{-(\mu_k - \alpha_r - \beta_{k'})}} - \frac{1}{1 + e^{-(\mu_{k-1} - \alpha_r - \beta_{k'})}},$$
$$r = 1, \ldots, R;\ k = 2 \ldots q;\ k' = 1 \ldots q;\ \theta_{rk'1} = \frac{1}{1 + e^{-(\mu_1 - \alpha_r - \beta_{k'})}}$$
$$\mu \mid \sigma_\mu^2 \overset{iid}{\sim} \text{OS}[\text{Normal}(0, \sigma_\mu^2)],\ \mu_k > \mu_{k-1}; k = 1 \ldots q;\ \mu_q = \infty$$
$$\alpha_r \mid \sigma_\alpha^2 \overset{iid}{\sim} \text{Normal}(0, \sigma_\alpha^2),\ r = 1, \ldots, R;\ \alpha_1 = 0$$
$$\beta_{k'} \mid \sigma_\beta^2 \overset{iid}{\sim} \text{Normal}(0, \sigma_\beta^2),\ k' = 1 \ldots q;\ \beta_q = 0$$
$$\sigma_\mu^2 \sim \text{Inverse Gamma}(a_\mu, b_\mu)$$
$$\sigma_\alpha^2 \sim \text{Inverse Gamma}(a_\alpha, b_\alpha)$$
$$\sigma_\beta^2 \sim \text{Inverse Gamma}(a_\beta, b_\beta)$$
$$\pi \sim \text{Dirichlet}(\phi),\ r = 1 \ldots R$$

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 $\phi = 1.5$. I is the identity matrix.

In words, we assign Truncated Normal priors for the cutoff points μ, Normal priors centered on zero and with an unknown variance for α and β, a Dirichlet prior for the mixing probabilities π, and Inverse Gamma priors for the unknown variances σ^2_μ, σ^2_α and σ^2_β. Figure 2 shows a graphical representation of the model.

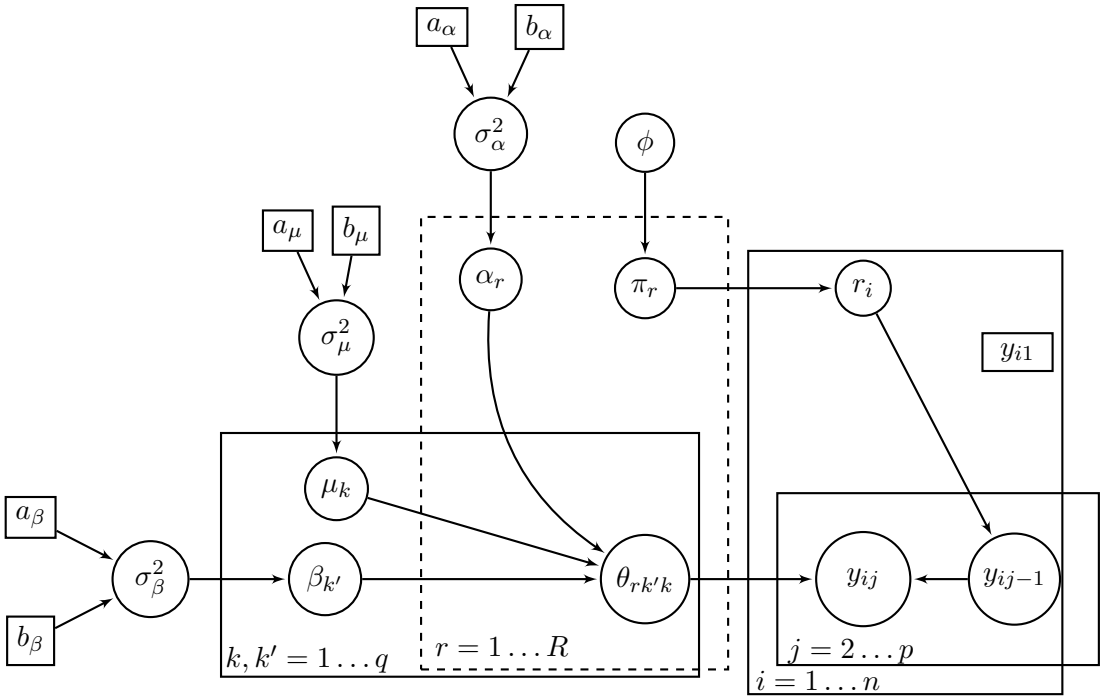


Figure 2: Graphical representation of the model

Given the likehood, equation 2, the posterior distributions for the model parameters are not available in close form. To perform the posterior computation, we use a Markov chain Monte Carlo (MCMC) sampling scheme. In particular, we use a Randon-Walk Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970) to sample block of parameters separately (μ, α, β and π and the parameters of the priors). For instance, to sample from the posterior of $\mu = (\mu_1, \dots, \mu_{q-1})$ we followed McKinley et al. (2015) and used the a truncated uniform with a fixed stepsize τ as a proposal. Specifically, we used the following algorithm

1. Set starting values for all the model parameters':
   $(\mu, \alpha, \beta, \pi, \sigma_\mu, \sigma_\alpha, \sigma_\beta) = (\mu_0, \alpha_0, \beta_0, \pi_0, \sigma_{\mu 0}, \sigma_{\alpha 0}, \sigma_{\beta 0})$

2. Set the stepsize of the proposal (τ)

3. Choose a μ_k for $k = 1, \dots, q-1$ at random and generate a new μ'_k candidate from its proposal

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

4. Accept μ'_k with probability

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

5. Repeat steps 3 and 4 until convergence.

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

3.3 Model Comparison

There are several ways to compare between models in a Bayesian framework: Bayes Factors (Kass & Raftery, 1995)), estimating the joint posterior distribution of all of the competing models using RJMCMC and others (Green, 1995; Richardson & Green, 1997) and using information criteria. We will use the latter approch here.

Importantly, (frequentist-like) information criteria that use a loss function evaluated at a point estimate can't be directly applied in a Bayesian setting. 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, also known as the label switching problem (Richardson & Green, 1997; Marin et al., 2005).

To compare among competing models we thus use the Widely Aplicable Information Criterion (WAIC) (Watanabe, 2009, 2010) that uses all the estimated posterior distribution. For a model with parameters ω and data Y, WAIC is defined as

$$\text{WAIC} = -2 \sum_{i=1}^{n} \log \int p(Y_i|\omega) p(\omega|Y) d(\omega) + 2p$$

Where:

- $p = \sum_{i=1}^{n} \{\log \int p(Y_i|\omega) p(\omega|Y) d(\omega) - \int log\, p(Y_i|\omega) p(\omega|Y) d(\omega)\}$ is the number of effective parameters

Alternatively, the number of effective parameters could also be approximated by $p_2 = \sum_{i=n}^{S} \text{Variance}[\log p(Y_i|\omega)]$. Defined this way the WAIC is in the same scale as other information criteria. Gelman et al. (2014) calls the observation i contribution to the likelihood $p(Y_i|\omega)$ 'log pointwise predictive density'. We follow this terminology here and call WAIC's first component 'log predictive density' (LPD).

As a comparison, we also present the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002; Spiegelhalter, Best, Carlin, & Linde, 2014) calculated using the relabelled MCMC chains using the algorithm proposed by Stephens (2000). We separate the two components of the DIC: Mean Deviance ($\bar{D}$) and number of effective parameters (p_d) so that these could be adequately compared with the WAIC components. The DIC is been used extensively in Bayesian applications, although it has the limitation of needing a point estimate that adequately represents the estimated posterior. If this is not the case, ie with multimodal posteriors, p_d could be negative. Table 3 (Appendix B) shows model comparison using unprocessed MCMC output and is a good example of this.

4. Results

We estimate the model using a random subsample of 442 individuals that had their SRHS recorded in all waves (about 10% of the total). We used the R statistical language (R 3.02) linked with C++ routines and fitted models where the number of latent groups (R) varying from one (R=1, no-clustering) to six (R=6, six latent groups or row-clusters). The MCMC

chain was ran for about a million iterations (3 parallel chains and around 200,000 burn-in) and post-processed the output according to the the algorithm of Stephens (2000) to rectify label switching. [2] For each fitted model, we present the following: number of clusters (R), total number of parameters (Pars), DIC ($\bar{D}$ and p_d) and the two versions of the WAIC (LPD, p, p_2). Table 2 shows the results. All the information criteria suggest the same conclusion, the model with five clusters seems to provide the best fit.

Table 2: Model Comparison relabelled chains

R	Pars	$\bar{D}$	p_d	DIC	LPD	p	WAIC	p_2	WAIC2
1	7	3564.8	6.8	3571.6	3555.9	8.9	3573.8	9.0	3573.9
2	10	3233.9	8.3	3242.1	3218.3	15.6	3249.4	15.7	3249.8
3	12	3073.5	9.9	3083.3	3061.6	11.8	3085.3	12.0	3085.5
4	14	3042.2	11.4	3053.6	3029.9	12.3	3054.5	12.5	3054.9
5	16	3035.2	12.9	3048.1	3022.5	12.8	3048.0	13.0	3048.4
6	18	3035.6	13.0	3048.6	3022.6	12.9	3048.5	13.2	3049.0

What do these estimated five row-clusters look like? Figure 3 shows estimated transition matrices for three groups (out of the five). The estimated probabilities of having the same response in the current period are given by the diagonal of the matrix. Averaged over 2001-2011, individuals tend to move towards the middle responses in their year to year transition. That is, individuals that whose response was "Poor" ("Excellent") are more likely to report "Fair" ("Very Good") the next period. The groups with negative cluster effect ($\alpha < 0$) are more likely to have responses in the end of the scale ("Poor" and "Fair"') and not move in the next year. The opposite is true for groups with a negative α. They are more likely to have responded "Very Good" or "Excellent" and have a similar response in the next period.

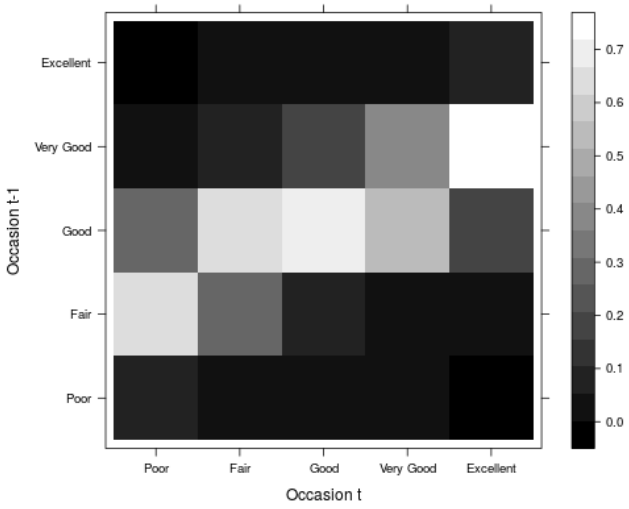
5. Discussion and Conclusions

Model-based cluster models provide a way to estimate latent groups and more generally reduce the data dimensionality even for correlated data. In this paper, we have used finite mixtures of cumulative logits that include the past response as a covariate to model repeated ordinal data. We estimated the model in a Bayesian setting using MCMC with a block Metropolis-Hastings sampler. To compare among models with different number of mixture components we used WAIC and DIC. Relabelling strategies allowed us to identify the latent groups itself, that is each mixture component.

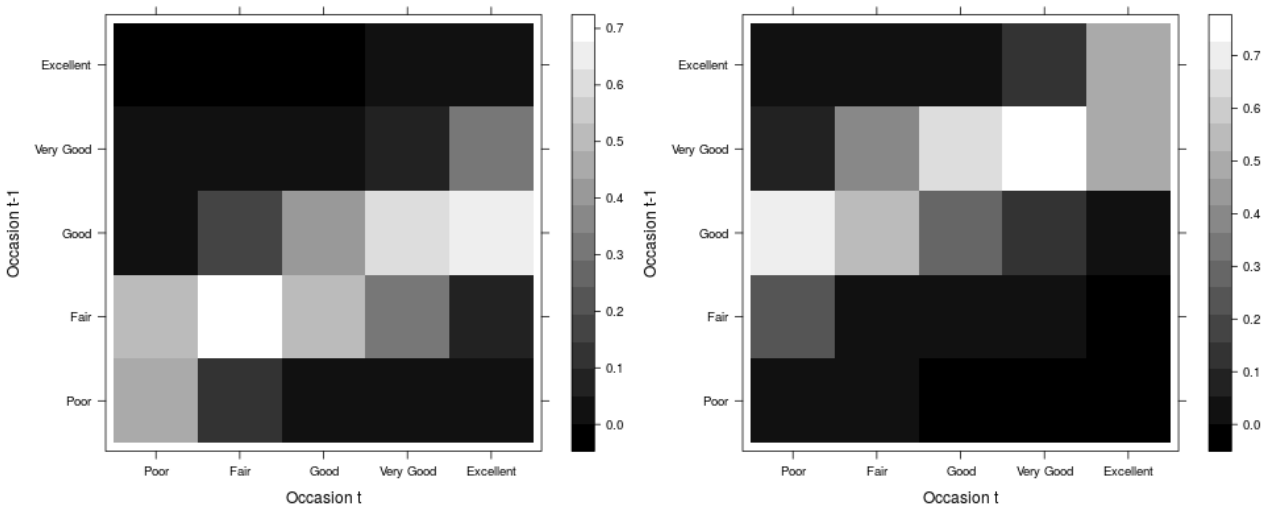
As an ilustration, we applied the model to self-reported health status data (poor, fair, good, very good and excellent) over 11 years in Australia and found evidence for five latent groups with distinct transitions overtime: two where it remains stable, two where it improves and one where it worsens.

The model has some limitations. Firstly, it is computer-intensive and estimation might become impractical with big datasets (hundred of thousands). 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 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.

[2] Results for the unprocessed MCMC, including traceplots that ilustrate the label switching problem could be found in Appendix A.



$\alpha_r = 0$ (baseline)



$\alpha_r < 0$ $\alpha_r > 0$

Figure 3: Transition matrices $\hat{\theta}_{rk'k}$ for several clusters

Secondly, caution should be taken on the interpretation of the number and in general individual mixture components. Mixture models are very flexible and with the enough number of components could fit any dataset. Information criteria like WAIC and DIC penalise model complexity but as the Bayesian equivalents of the AIC they potentially also select too many cluster components. Measures like the WBIC (Watanabe, 2013) that include a bigger penalty for model complexity could be worth exploring here.

Lastly, Bayesian approaches can always be sensitive to choice of priors. Our weakly

informative priors could even make this problem worse. In order to rule out this, we have ran simulations with different priors and similar sample size and number of cluster to the SRHS data. Although not shown here we found that the conclusions are robust to the choice of priors. A simulation study could nonetheless be important for a more comprehensive check.

In addition to the above, we plan to extend the model in two directions: exploring other ways to incorporate the correlation and including the number of mixture components as parameter 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 trans-dimensional models such as RJMCMC (Green, 1995; Richardson & Green, 1997) or Bayesian Non-Parametric models (Müller, Quintana, Jara, & Hanson, 2015). Albeit more complex these models have also the advantage of estimating a posterior for the number of mixture components and thus simplify the comparison of competing models.

References

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

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.

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

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*.

Fernández, D., Arnold, R., & Pledger, S. (2014). Mixture-based clustering for the ordered stereotype model. *Computational Statistics & Data Analysis*.

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

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

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

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, *82*(4), 711–732.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

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

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

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.

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

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

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

McKinley, T. J., Morters, M., Wood, J. L., 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, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

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

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

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

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

Robert, C. P., & Casella, G. (2005). *Monte carlo statistical methods (springer texts in statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

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

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

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

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.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*. Cambridge University Press.

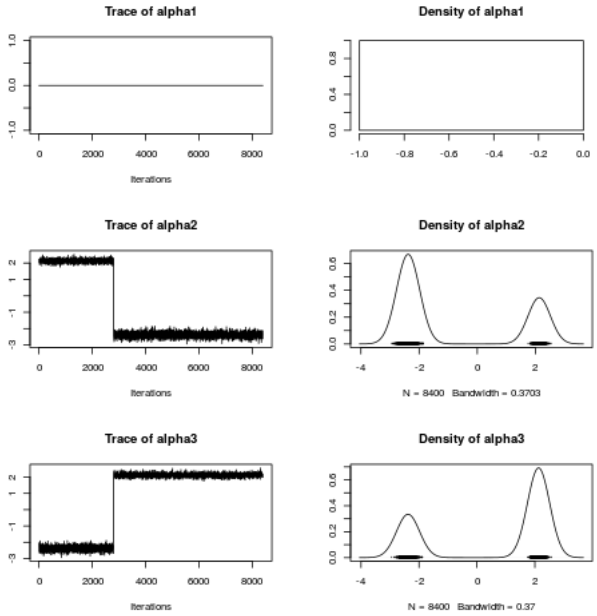
Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

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

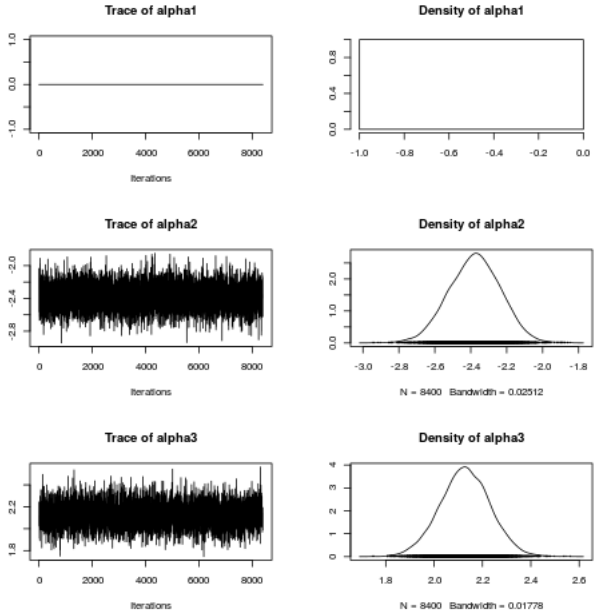
Appendix

A. Unprocessed and relabelled MCMC output for selected parameters

Unprocessed (selected α's)



Relabelled (selected α's)



B. DIC estimates using unprocessed chains

Table 3: Model comparison unprocessed chains

R	Pars	$\bar{D}$	p_d	DIC	LPD	p	WAIC	p_2	WAIC2
1	7	3564.8	6.8	3571.6	3555.9	8.9	3573.8	9.0	3573.9
2	10	3233.9	-703.1	2530.8	3218.3	15.6	3249.4	15.7	3249.8
3	12	3073.5	-335.0	2738.4	3061.6	11.8	3085.3	12.0	3085.5
4	14	3042.2	-442.5	2599.7	3029.9	12.3	3054.5	12.5	3054.9
5	16	3035.2	-131.4	2903.8	3022.5	12.8	3048.0	13.0	3048.4
6	18	3035.6	-508.7	2526.9	3022.6	12.9	3048.5	13.2	3049.0

C. Proposals

Choose initial values for all the parameters in the model μ, α, β, π, σ^2_μ, σ^2_α, and σ^2_β, and update them according to the following:

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

$$\alpha'_r \mid \alpha_r \overset{iid}{\sim} \text{Normal}(\alpha_r, \sigma^2_{\alpha p}) \qquad r = 2 \ldots R,\ \alpha_1 = 0$$

$$\beta'_j \mid \beta_j \overset{iid}{\sim} \text{Normal}(\beta_j, \sigma^2_{\beta p}) \qquad j = 2 \ldots p,\ \beta_1 = 0$$

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

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

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

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

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

Where the proposals "steps" τ, $\sigma^2_{\alpha p}$, $\sigma^2_{\beta p}$, $\sigma^2_{\pi p}$, $\sigma^2_{\sigma\mu p}$, $\sigma^2_{\sigma\alpha p}$ and $\sigma^2_{\sigma\beta p}$ are fixed.