Clustering Ordinal Data via Latent Variable Models

Damien McParland and Isobel Claire Gormley

Abstract Item response modelling is a well established method for analysing ordinal response data. Ordinal data are typically collected as responses to a number of questions or items. The observed data can be viewed as discrete versions of an underlying latent Gaussian variable. Item response models assume that this latent variable (and therefore the observed ordinal response) is a function of both respondent specific and item specific parameters. However, item response models assume a homogeneous population in that the item specific parameters are assumed to be the same for all respondents. Often a population is heterogeneous and clusters of respondents exist; members of different clusters may view the items differently. A mixture of item response models is developed to provide clustering capabilities in the context of ordinal response data. The model is estimated within the Bayesian paradigm and is illustrated through an application to an ordinal response data set resulting from a clinical trial involving self-assessment of arthritis.

1 Introduction

Ordinal data arise naturally in many different fields and are typically collected as responses to a number of questions or items. A common approach to analysing such data is to view the observed ordinal data as discrete versions of an underlying latent Gaussian 'generating' variable. Many models such as graded response models and ordinal regression models [2] make use of this concept of latent generating variables.

Item response modelling [4] is an established method for analysing ordinal response data. It is assumed that the observed ordinal response to an item will be level k, say, if the underlying latent variable lies within a specified interval. Item response models further assume that the latent generating variable (and therefore the

Damien McParland and Isobel Claire Gormley
University College Dublin, Ireland.
e-mail: {damien.mcparland,claire.gormley}@ucd.ie

observed ordinal response) is a function of both respondent specific and item specific parameters. The respondent specific parameters are often called *latent traits*. The probability of a certain response from a respondent is related to both the value of their latent trait and also some item specific parameters.

Item response models assume that the item specific parameters are the same for all respondents, i.e. a homogeneous population is assumed. Often a population is heterogeneous however and clusters of respondents exist; members of different clusters may view the items differently. Here an item response model is embedded in a mixture modelling framework to facilitate clustering of respondents in the context of ordinal response data. Under the mixture of item response models the probability that a respondent gives a certain response depends on their latent trait and on group specific item parameters. An alternative approach to this problem is given in [17].

The mixture of item response models is developed and estimated within the Bayesian paradigm using Markov chain Monte Carlo methods. A key issue is choosing the optimal model or equivalently, the number of components in the optimal mixture model. The marginal likelihood is employed here to choose between models and a bridge sampling approach to estimating the marginal likelihood is used.

The model is illustrated through an application to an ordinal response data set resulting from a clinical trial involving self-assessment of arthritis pain levels.

The paper proceeds as follows. In Section 2 the arthritis pain levels data set used to demonstrate the model is introduced. Item response models and their extension to a mixture of item response models are considered in Section 3. Section 4 is concerned with Bayesian model estimation and also model selection. The results from fitting the model to the illustrative data set are presented in Section 5. Finally, discussion of the model takes place in Section 6.

2 Arthritis pain data

An ordinal data set is employed to illustrate the mixture of item response models. The data come from a clinical trial in which patients suffering from rheumatoid arthritis are randomly assigned to a treatment group or a placebo group. The patients self-assess their arthritis related pain as 1 (poor), 2 (fair) or 3 (good) at one and five month examinations. Some covariate information associated with each patient such as their age and sex are also recorded. Further details are given in [12] and [1].

Here only the ordinal response data are analysed. Interest lies in determining if there is an underlying group structure among the group of 289 patients in the clinical trial. Members of the same group would be expected to have similar arthritis pain profiles. In particular, whether or not patients in the treatment group are differentiated from the patients in the placebo group is of interest.

3 Item Response Models and Mixtures of Item Response Models

The concepts behind item response models and the proposed extension to a mixture of item response models are explained in this section.

3.1 Item Response Models for Ordinal Data

Suppose the responses of N individuals to each of J items are observed. Since the data are ordinal, the set of possible responses to item j is $\{1,2,\ldots,K_j\}$ where K_j denotes the number of possible responses to item j. Thus the data can be represented by an $N \times J$ matrix, Y, where y_{ij} is the response of individual i to item j.

Corresponding to each ordinal response, y_{ij}, is a latent Gaussian variable, z_{ij}. A Gaussian link function is used here but other link functions, such as the logit [4], can be employed. For each item there exists a vector of threshold parameters $\underline{\gamma}_j = (\gamma_{j,0}, \gamma_{j,1}, \ldots, \gamma_{j,K_j})$. This vector is subject to the constraint:

$$-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq \ldots \leq \gamma_{j,K_j} = \infty$$

The observed ordinal response, y_{ij}, serves as an indicator to the latent variable z_{ij}:

$$y_{ij} = k \;\Rightarrow\; \gamma_{j,k-1} \leq z_{ij} \leq \gamma_{j,k}$$

In addition to the latent variable, z_{ij}, it is assumed that there exists a latent trait vector, $\underline{\theta}_i$, of dimension q corresponding to each individual. Here q is user specified. The mean of the conditional distribution of z_{ij} is related to this latent trait:

$$z_{ij}|\underline{\theta}_i \sim N(\underline{\lambda}_j^T \underline{\theta}_i - b_j, 1)$$

In the item response literature the item parameters $\underline{\lambda}_j$ and b_j are usually termed *item discrimination* parameters and *item difficulty* parameters respectively. The conditional probability that a response takes a certain ordinal value can then be expressed as the difference between two standard Gaussian cumulative density functions:

$$P(y_{ij} = k|\underline{\lambda}_j, b_j, \underline{\gamma}_j, \underline{\theta}_i) = \Phi[\gamma_{j,k} - (\underline{\lambda}_j^T \underline{\theta}_i - b_j)] - \Phi[\gamma_{j,k-1} - (\underline{\lambda}_j^T \underline{\theta}_i - b_j)]$$

3.2 A Mixture of Item Response Models (MIRM) for Ordinal Data

A mixture modelling framework can be imposed on the item response model for cases where there is an underlying group structure in the data. The aim of this mixture of item response models is to cluster individuals into their unobservable groups. Under the MIRM, the latent variable z_{ij} is a mixture of G Gaussian densities:

$$f(z_{ij}) = \sum_{g=1}^{G} \pi_g N(\underline{\lambda}_{gj}^T \underline{\theta}_i - b_{gj}, 1)$$

The probability of belonging to group g is denoted π_g while $\underline{\lambda}_{gj}$ and b_{gj} are *group specific* item discrimination and difficulty parameters respectively.

A latent indicator variable, $\underline{\ell}_i = (\ell_{i1}, \ldots, \ell_{iG})$ is introduced for each individual i. This binary vector indicates to which group individual i belongs i.e. $l_{ig} = 1$ if i belongs to group g; all other entries in the vector are 0. Thus, conditional on $\underline{\ell}_i$, the probability of observing a particular ordinal response is:

$$P(y_{ij} = k | \underline{\lambda}_{gj}, b_{gj}, \underline{\gamma}_j, \underline{\theta}_i, l_{ig} = 1) = \Phi\left[\gamma_{jk} - (\underline{\lambda}_{gj}^T \underline{\theta}_i - b_{gj})\right] - \Phi\left[\gamma_{j,k-1} - (\underline{\lambda}_{gj}^T \underline{\theta}_i - b_{gj})\right]$$

The augmented likelihood, $\mathscr{L}(\Lambda, B, \Gamma, \Theta, L, Z|Y)$, is given by:

$$\prod_{i=1}^{N} \prod_{g=1}^{G} \prod_{j=1}^{J} \left\{ \left[\sum_{k=1}^{K_j} \mathbf{1}\left(\gamma_{j,k-1} \leq z_{ij} \leq \gamma_{j,k}\right) \mathbf{1}\left(y_{ij} = k\right) \right] N\left(\underline{\lambda}_{gj}^T \underline{\theta}_i - b_{gj}, 1\right) \right\}^{\ell_{ig}}$$

An assumption of local independence is implicit here, i.e. conditional on the latent trait $\underline{\theta}_i$ the J responses by individual i are independent. The responses of different individuals are also regarded as independent.

4 Parameter Estimation and Model Selection

The Bayesian framework in which the model is estimated, the Markov chain Monte Carlo (MCMC) algorithm used to fit the model and the bridge sampling algorithm which facilitates model selection are all described in what follows.

4.1 Prior and Posterior Distributions

To implement the model described above in a Bayesian framework prior distributions must be specified for all unknown parameters. Priors are required for the threshold parameters $\underline{\gamma}_j$, the item parameters, $\underline{\lambda}_{gj}$ and $\underline{b}_g$, and for the mixing weights $\underline{\pi}$ (for $j = 1, \ldots, J$ and $g = 1, \ldots, G$). Specifically, a uniform prior is specified for the threshold parameters and for the other parameters the prior distributions are:

$$p(\underline{\lambda}_{gj}) = MVN_q(\underline{\mu}_\lambda, \Sigma_\lambda) \qquad p(\underline{b}_g) = MVN_J(\underline{\mu}_b, s_b^2 \mathbf{I}) \qquad p(\underline{\pi}) = Dir(\underline{\alpha})$$

The posterior distribution is:

$$p(\Lambda, B, \Gamma, \underline{\pi}, \Theta, L, Z|Y) \propto \mathscr{L}(\Lambda, B, \Gamma, \Theta, L, Z|Y) p(\Lambda) p(B) p(\Gamma) p(\Theta) p(L|\underline{\pi}) p(\underline{\pi})$$

where $p(\Lambda)$, $p(B)$, $p(\Gamma)$ and $p(\underline{\pi})$ are the prior distributions detailed above. The latent trait variable $\underline{\theta}_i$ is assumed to have a standard multivariate Gaussian distribution; the latent indicator variables $\underline{l}_i$ follow a $Multinomial(1,\underline{\pi})$ distribution.

This model suffers from non-identifiability. To identify the model (as in [11]) the second element of each of the threshold vectors, $\underline{\gamma}_j$ for $j = 1,\ldots,J$, is fixed at 0. The model is also rotationally invariant. Therefore, a specific form is imposed on each matrix of discrimination parameters Λ_g for $g = 1,\ldots,G$. As in [8], the first q rows of this matrix are constrained to have a lower triangular form. In what follows, the free and fixed elements of the j^{th} row of Λ_g are denoted by $\underline{\lambda}^{\circ}_{gj}$ and $\underline{\lambda}^{\bullet}_{gj}$ respectively.

4.2 Estimation via a Markov Chain Monte Carlo Algorithm

The marginal distributions of the unknown parameters cannot be obtained analytically for this model so a MCMC algorithm is used to produce estimates of the model parameters. The algorithm used here is similar to the algorithm proposed in [3]. A Gibbs sampler is used to sample all latent variables and parameters, except the threshold parameters, $\underline{\gamma}_j$. These are sampled using a Metropolis-Hastings step.

Full conditional distributions for the model parameters and latent variables are:

- $\underline{\ell}_i|\ldots \sim \text{Multinomial}(1,\underline{p} = (p_1,\ldots,p_G))$ where

$$p_g \propto \pi_g \prod_{j=1}^{J}\left[\sum_{k=1}^{K_j} \mathbf{1}\left(\gamma_{j,k-1} \leq z_{ij} \leq \gamma_{j,k}\right)\mathbf{1}(y_{ij}=k)\right] \text{N}(\underline{\lambda}_{gj}^T\underline{\theta}_i - b_{gj}, 1)$$

- $\underline{\pi}|\ldots \sim \text{Dirichlet}(n_1+\alpha_1,\ldots,n_G+\alpha_G)$ where $n_g = \sum_{i=1}^{N} l_{ig}$.
- $z_{ij}|\ldots \sim \text{N}^T\left(\underline{\lambda}_{gj}^T\underline{\theta}_i - b_{gj}, 1\right)$ where the distribution is truncated to $[\gamma_{j,(y_{ij}-1)}, \gamma_{j,(y_{ij})}]$.
- $\underline{\theta}_i|\ldots \sim \text{MVN}_q\left[D_g^{-1}\Lambda_g^T\left(\underline{z}_i + \underline{b}_g\right), D_g^{-1}\right]$ where, $\underline{z}_i = (z_{i1},\ldots,z_{iJ})^T$ and $D_g = \Lambda_g^T\Lambda_g + \mathbf{I}_q$.
- $\underline{\lambda}^{\circ}_{gj}|\ldots \sim \text{MVN}\left\{S^{-1}\left[\Theta_g^{\circ T}\left(\underline{z}_{gj} - \Theta_g^{\bullet}\underline{\lambda}^{\bullet}_{gj} + b_{gj}\underline{1}\right) + \Sigma_{\lambda}^{-1}\underline{\mu}_{\lambda}\right], S^{-1}\right\}$ where $S = \left[\Sigma_{\lambda}^{-1} + \Theta_g^{\circ T}\Theta_g^{\circ}\right]$ and $\underline{1} = (1,\ldots,1)^T$. The i^{th} row of Θ_g° consists of the elements of $\underline{\theta}_i$ which multiply $\underline{\lambda}^{\circ}_{gj}$ for all individuals i in group g. Similarly $\Theta_g^{\bullet}$ consists of the elements which multiply $\underline{\lambda}^{\bullet}_{gj}$. The elements of the j^{th} column of the $N \times J$ matrix Z corresponding to individuals in group g are denoted by $\underline{z}_{gj}$.
- $b_{gj}|\ldots \sim N\left[(n_g + s_b^{-2})^{-1}(\underline{1}^T\Theta_g\underline{\lambda}_{gj} + s_b^{-2}\mu_{bj} - \underline{z}_{gj}^T\underline{1}), (n_g + s_b^{-2})^{-1}\right]$ where the rows of Θ_g are the latent trait vectors $\underline{\theta}_i$ for all individuals i in group g.

The posterior full conditional distribution of each of the threshold parameters, $\gamma_{j,k}$ can be shown to be uniform. When there are a large number of observations in adjacent categories this interval tends to be small which results in minimal movement of the Gibbs sampler. The algorithm therefore converges slowly. This difficulty is overcome by sampling from the posterior of the threshold parameters us-

ing a Metropolis-Hastings step, as in [3, 11]. Candidate values $v_{j,k}$ are proposed for $\gamma_{j,k}$ from the Gaussian distribution $\mathrm{N}^T(\gamma_{j,k}^{(t-1)}, \sigma_{MH}^2)$, truncated to the interval $(v_{j,k-1}, \gamma_{j,k+1}^{(t-1)})$ where $\gamma_{j,k+1}^{(t-1)}$ is the value of $\gamma_{j,k+1}$ at iteration $(t-1)$. The tuning parameter σ_{MH}^2 is selected to achieve appropriate acceptance rates.

4.3 Model Selection via the Bridge Sampler

Since the proposed model is a finite mixture model, the number of components G in the mixture must be chosen. A bridge sampling algorithm [16, 6] is employed to approximate the marginal likelihood of a G component model. The marginal likelihood is evaluated for a range of models with different values of G and the model with the highest marginal likelihood is chosen as optimal. Here, the posterior mean of the latent Gaussian variable Z is treated as the 'observed data'. This approach removes the need to work with the intractable marginal distribution of the ordinal data, Y, and also the posterior distribution of the threshold parameters.

In order to use bridge sampling to approximate the marginal likelihood it is important that the MCMC algorithm mixes well over all posterior modes. The random permutation MCMC sampler [5] is used to achieve this. For more details on the bridge sampling estimator of the marginal likelihood of a mixture model see [7].

5 Arthritis pain data: results.

The mixture of item response models (MIRM) was fitted to the ordinal arthritis pain data described in Section 2. A number of mixture of item response models were fitted to the data with the number of components G ranging from one to five, and with a user specified $q = 1$ dimensional latent trait. The marginal likelihood of each of the models was estimated using the bridge sampling technique described in Section 4.3. The values obtained are illustrated in Figure 1. The highest marginal likelihood value is obtained when a two component MIRM is fitted. Posterior mean parameter estimates for the optimal model are detailed in Table 1.

Inspection of the responses of individuals in each cluster suggests that the patients have been partitioned into a group (group 1) who judge the state of their arthritis to be poor to fair and a group (group 2) who consider the state of their arthritis to be fair to good. Although the item difficulty parameters for both groups are negative, the parameters for group 1 [$\underline{b}_1 = (-0.18, -0.20)$] are smaller in magnitude than those for group 2 [$\underline{b}_2 = (-2.29, -2.27)$]. This difference means that the values of the latent Gaussian variable Z (with marginal mean $-\underline{b}_g$ for $g = 1, 2$) are lower in group 1, reflecting the generally lower observed ordinal responses found in group 1. The confidence regions for the discrimination parameters include 0 which indicates that even the one dimensional latent trait may be unnecessary for this data

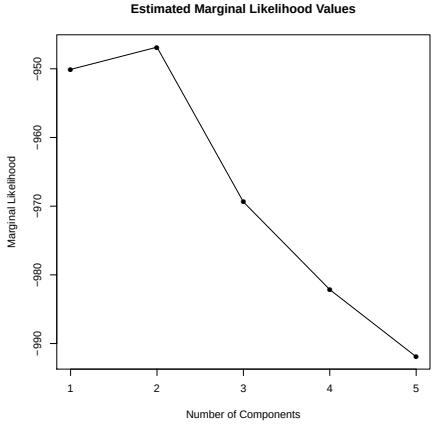

Fig. 1: Estimated marginal likelihood values for a range of mixture of item response models with a one dimensional latent trait.

Parameter	Posterior mean	
b_{11}	-0.18	[-1.02, 0.47]
b_{12}	-0.20	[-1.11, 0.70]
b_{21}	-2.29	[-3.35, -1.50]
b_{22}	-2.27	[-3.35, -1.44]
λ_{11}	0.59	[-0.25, 1.49]
λ_{12}	0.81	[-0.32, 1.71]
λ_{21}	0.96	[-0.02, 1.79]
λ_{22}	0.75	[-0.10, 1.54]
$\gamma_{1,2}$	2.10	[1.59, 2.97]
$\gamma_{2,2}$	1.78	[1.30, 2.63]
π_1	0.40	[0.22, 0.60]
π_2	0.60	[0.40, 0.78]

Table 1: Posterior mean estimates (and 95% quantile-based confidence regions) for the optimal model.

set. Interestingly, the two groups uncovered by the model do not correspond to the treatment and placebo group (Rand index = 0.51, Adjusted Rand index = 0.015).

6 Discussion

Ordinal data arise in many different fields. The mixture of item response models presented here facilitates the clustering of such data. This is achieved by assuming the observed ordinal data are discrete versions of an underlying latent Gaussian variable. The clustering is achieved by fitting a mixture model to the latent Gaussian data. The model is closely related to the mixture of factor analysers model [14, 15] for continuous data; in the case of the mixture of item response models however, only a discrete version of the data are observed.

Bridge sampling was employed for model selection. Simulation studies and the illustrative data example suggest that the bridge sampling approach works well in the context of the mixture of item response models. However, it should be noted that as the bridge sampler relies on the posterior mean of the latent Gaussian data Z, the same 'data' are not used when evaluating the marginal likelihood for different models. Again, simulation studies suggest that given a sufficiently large data set (both in terms of number of observations and cell counts for the ordinal variables) the results are not very sensitive to this approximation to Y.

There are a number of ways in which the model could be extended. The model selection technique employed here is used only to choose the number of compo-

nents in the mixture. Extending the bridge sampling technique to determine the optimal number of dimensions (q) for the latent trait would be very beneficial [13]. Additionally, in the illustrative data set used here covariate data were available. Incorporating these data in the model would be potentially informative and could be achieved within a mixture of experts framework [10, 9]. Finally, as with most clustering models, the set of variables on which the clustering is based strongly influences the MIRM; incorporating a variable selection step while clustering would potentially improve clustering performance.

7 Acknowledgements

This work has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 09/RFP/MTH2367.

References

1. Agresti, A. Analysis of Ordinal Categorical Data. Wiley. (2010)
2. Albert, J.H. and Chib, S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association. 88:669–679. (1993)
3. Cowles, M.K. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. Journal of the American Statistical Association. 6:101–111. (1996)
4. Fox, J. P. Bayesian Item Response Modeling. Springer. (2010)
5. Frühwirth-Schnatter, S. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association. 96:194–209. (2001)
6. Frühwirth-Schnatter, S. Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. Statistica Sinica. 6:831–860. (2004)
7. Frühwirth-Schnatter, S. Finite Mixture and Markov Switching Models. Springer. (2006)
8. Geweke, J. and Zhou, G. Measuring the price of arbitrage theory. The Review of Financial Studies. 9:557–587. (1996)
9. Gormley, I.C. and Murphy, T.B. A mixture of experts model for rank data with applications in election studies. The Annals of Applied Statistics, 2(4), 1452–1477. (2008)
10. Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixture of local experts. Neural Computation, 3, 79–87. (1991)
11. Johnson, V. E. and Albert, J. H. Ordinal Data Modeling. Springer, New York. (1999)
12. Lipsitz, S. R. and Zhao, L. Analysis of repeated categorical data using generalized estimating equations. Statistics in Medicine. 13, 1149–1163. (1994)
13. Lopes, H. F. and West, M. Bayesian model assessment in factor analysis. Statistica Sinica 14, 41–67. (2004)
14. McLachlan, G. J. and Peel, D. Finite mixture models, John Wiley & Sons, New York. (2000)
15. McNicholas, P.D. and Murphy, T. B. Parsimonious Gaussian mixture models. Statistics and Computing, 18(3), 285–296. (2008)
16. Meng, X. L. and Wong, W. H. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. The Econometrics Journal. 7:143–167. (1996)
17. Von Davier, M. and Yamamoto, K. Partially observed mixtures of IRT Models:An extension of the generalized partial credit model. Applied Psychological Measurement, 28(6), 389–406. (2004)