A full Bayesian implementation of a generalised partial credit model with an application to an international disability survey

Sujit K. Sahu[1], Mark R. Bass[2], Carla Sabariego[3],
Alarcos Cieza[4], Carolina S. Fellinghauer[5], Somnath Chatterji[6] *

September 16, 2019
To appear in the *Journal of the Royal Statistical Society*, Series C, Applied Statistics.

Abstract

Generalised partial credit models (GPCM) are ubiquitous in many applications in the health and medical sciences that use item response theory. Such polytomous item response models have a great many uses ranging from assessing and predicting an individual's latent trait to ordering the items to test the effectiveness of the test instrumentation. By implementing these models in a full Bayesian framework, computed through the use of Markov chain Monte Carlo (MCMC) methods implemented in the efficient STAN software package, this article exploits the full inferential capability of the GPCMs. The GPCMs include explanatory covariate effects which allow simultaneous estimation of regression and item parameters. The Bayesian methods for ranking the items using the Fisher information criterion (FIC) are implemented using MCMC. This allows us to fully propagate and ascertain uncertainty in the inferences by calculating the posterior predictive distribution of item specific FIC in a novel manner that has not been exploited in the literature before. Lastly, we propose a new Monte Carlo method for predicting the latent trait score of a new individual by approximating the relevant Bayesian predictive distribution. Data from a Model Disability Survey carried out in Sri Lanka by the World Health Organisation (WHO) and the World Bank are used to illustrate the methods. The proposed approaches are shown to provide simultaneous model based inference for all aspects of disability which can be explained by environmental and socio-economic factors.

Keywords: Bayesian Methods, Education Testing, Hierarchical Modelling, Item ranking, Item Response Theory.

1 Introduction

Generalised partial credit models, proposed by Muraki (1992), for modelling polytomous response data are applied in diverse fields and applications such as health, Li and Baser (2012); Verhagen and Fox

*1: **Corresponding Author**, Email: *S.K.Sahu@soton.ac.uk*, School of Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK; 2: Barclays Services Limited, London, UK; 3: Jointly appointed at (1) Disability Policy and Implementation Research Group, Swiss Paraplegic Research, Nottwil, Switzerland and (2) Department of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland; 4: Department for Management of Noncommunicable Diseases, Disability, Violence and Injury Prevention, World Health Organization, Geneva 1211, Switzerland; 5: Swiss Paraplegic Research, Nottwil 6207, Switzerland; 6: Department of Information, Evidence and Research, World Health Organization, Geneva 1211, Switzerland.

(2013) and educational progress testing, Patz and Junker (1999). These models are examples in the generic field of item response theory (IRT) for which there is an extensive literature (Birnbaum, 1968; Rasch, 1961; Samejima, 1969; Lord, 1980; Bock and Aitkin, 1981). Increasing research interests in the GPCMs started with the publication of the article by Muraki (1992) where the EM algorithm was proposed to fit these models; see e.g. Falk and Cai (2016) for recent developments in this field. By now there is a rich variety of fitting techniques and algorithms that facilitate inference using the GPCMs.

Development of Bayesian methodologies, see e.g. Mislevy (1986); Swaminathan and Gifford (1986); Tsutakawa and Lin (1986) for IRT has led to the development of IRT specific software packages, such as the `MultiLCIRT` by Bartolucci et al. (2014) which is an R (R Core Team, 2016) package that can fit IRT models for binary and ordinal polytomous items. The adoption of the powerful MCMC techniques for fitting item response models (see e.g. Albert (1992); Patz and Junker (1999); Fox and Glas (2001); Sahu (2002); Glas and Meijer (2003); Sinharay et al. (2006); Fox (2010); Li and Baser (2012); Johnson and Kuhn (2015)) has enabled the provision for richness in model checking and inference. Software packages such as `ltm` (Rizopoulos, 2006), `gpcm` (Johnson, 2007) and `ltbayes` (Johnson and Kuhn, 2015) also allow model fitting and estimation.

Recently, the open source Bayesian software package `STAN`, Stan Development Team (2015), is beginning to be used, see e.g. Furr et al. (2016). Software packages written in the R language together with model fitting using the general purpose Gibbs sampling software `WinBUGS` and `OpenBUGS` are also increasingly being used, see for example, Curtis (2010); Li and Baser (2012); Thomas et al. (2006). However, there is still a lack of literature on exploiting the rich inferential capabilities of a full Bayesian GPCM. To address this gap in the literature this article develops Bayesian computation methods to achieve three important inferential tasks, all of which are motivated below by a practical problem of analysing data from a comprehensive international disability survey developed by the WHO and the World Bank.

The Model Disability Survey (MDS)[1], developed as part of WHO's Global Disability Action Plan[2], is a standardised instrument for data collection on disability that provides comprehensive and systematic documentation on all aspects of health and functioning in a population. It is a general population survey, measuring how persons with different levels of disability conduct their lives while identifying the hindering and facilitating aspects of their environments. The MDS makes it possible for countries to collect information, not merely about persons who experience very significant levels of disability, but about those along the entire disability spectrum ranging from no disability to complete disability. This is especially important as it allows countries to develop public health strategies and policy interventions that promote, maintain and enhance functioning for people with varying degrees of disability. The MDS, used alone or as a module in a larger population survey, enables countries to collect internationally comparable disability data for national health and social policy planning purposes. Furthermore, it can be used to monitor the implementation of the requirements of the United Nations Convention of the Rights of Persons with Disabilities.[3]

MDS, being a very large scale comprehensive survey, collects information on a large (more than 100) number of items describing disability such as intrinsic capacity, daily functioning and environmental factors and also many socio-economic factors such as age, gender and annual income, see e.g. Sabariego et al. (2015). To estimate the effects of the socio-economic factors these authors propose a non-Bayesian approach of regressing the estimated latent trait scores, measuring physical and psychological disability,

[1] http://www.who.int/disabilities/data/mds/en/

[2] http://www.who.int/disabilities/actionplan/en/

[3] http://www.un.org/disabilities/convention/conventionfull.shtml

for capacity and performance, which are obtained using a Rasch model analysis. This approach, though useful in practice, does not correctly assess the uncertainties of the regression parameter estimates since it does not allow for uncertainties present in the estimated latent trait scores. At the modelling stage the GPCM is allowed to have any number of explanatory factors, e.g. environmental and socio-economic. Although there is a large literature on using the GPCM, only a handful articles incorporate simultaneous estimation of the covariate effects and the parameters of an IRT model, see e.g. Furr et al. (2016); Karabatsos (2017). The approach adopted in this paper is similar in spirit to the explanatory item response models collected in the edited book, de Boeck and Wilson (2004) mostly from a likelihood based inference point of view, see e.g. Glas (2005), and not specifically for GPCM.

The second inferential objective comes from the need to develop evidence-based brief versions of the core modules in the MDS that preserves their ability to capture essential information but can be more easily implemented in any national data collection platform or population survey. This article develops statistical methodology for creating this brief version of the MDS by ranking the importance of each item to explain the total information content in metrical scales (or latent trait scores). In this article the total information content is defined as the sum total of the expected Fisher information criterion (FIC) over all individuals for each item, where the expectation is evaluated with respect to the posterior distribution of the unknown parameters. Although the Fisher Information is defined in the same way as in the literature, see e.g. Muraki (1993) and Li and Baser (2012), our proposal for ranking the items differs fundamentally since these articles base their comparisons on the item characteristic curve obtained with plug-in estimates of the item specific parameters, see e.g. Ramsay (1991), whereas we use the entire Bayesian predictive distribution which takes care of the associated uncertainty.

The final inferential task is motivated by the need to predict the latent trait scores of new individuals who are not included in the current survey. This helps to achieve several objectives. For example, it eliminates the need for re-fitting the model when data from new individuals are available at a later date. Using the methods developed here it is possible to predict the latent trait scores of the new individuals which can be used to monitor latent trait over a longer time period. The methodology developed here is similar to that of Li and Lissitz (2004). However, the primary goal of their article was to derive analytical expressions for the standard errors of item parameter estimates in a classical inference case, while the contribution of this article lies in developing a predictive computational tool for estimating the latent trait score and not the item parameters. Moreover, the proposed method is designed to be implemented after MCMC model fitting in a Bayesian inference framework which has not been attempted in the literature as far as we are aware.

The remainder of this article is organised as follows. The motivating MDS data set is described in Section 2. Section 3 details the Bayesian modelling developments. Model fitting and prediction results for a selected data set from the MDS illustrate the Bayesian methods in Section 5. A few summary remarks are provided in Section 6. An online supplement contains further results with larger tables and graphs and also the data set used and the STAN code developed here to reproduce the numerical results of this paper.

2 Descriptive analysis and data exploration

We have data from a national survey conducted in Sri Lanka which was financed by the World Bank. One of the aims of this survey was to develop a continuous measure of latent trait across a representative sample of the general population and derive the most parsimonious set of items to measure this trait in a cross-population comparable manner. The survey was completed by $n = 3000$ individuals, of which

1791 were females (59.7%), each responding to $J = 17$ items that mainly relate to an individual's capacity to perform certain tasks. For each item, respondents were asked to select one of $K = 5$ categories, ranging from 1, "no difficulty" to 5, "extreme difficulty" in capacity levels. Table 5 in the online Supplement provides a brief description of the items along with the total number of respondents for each category for each item. Overwhelmingly, people responded in category one which implies good capacity levels.

Values of three covariates: age, gender and household income (hereafter referred to as income) were also recorded for each individual. Age, which varied between 17 and 96, is treated as a continuous covariate in our modelling. The respondents' income varied between 0 and 500,000 in the local currency, the Sri Lankan Rupee (LKR). This large range of the income distribution makes it problematic to include it as a continuous explanatory variable. To tackle this we treat income as an explanatory factor having five levels which are obtained by using the four income quintiles, 18, 25, 35, and 50 (in thousands) LKR of the income distribution. The mean incomes in the five groups were 10,651, 20,063, 27,949, 39,083 and 71,524 LKR.

The total score for each individual from the 17 items can vary between 17, corresponding to the best health state to 85, corresponding to the worst possible health state. The left panel of Figure 1 provides separate boxplot of the total scores for the males and females. The median total scores for the males and females in this figure are 19 and 20 respectively; the corresponding means are 23.5 and 24.7. Thus, the males, on average, report lower levels of disability than the females. The difference in scores between the males and females, though very small, can still lead to significant gender effect since the sample sizes, 1209 for males and 1791 for the females, are large. The plots in the middle panel confirms the gap in self reported disability levels between the two genders widens with increasing age. This panel, however, does not indicate a very strong age and gender interaction effect, i.e. different slopes for males and females for the covariate age. Hence, we do not consider such interaction effects in the modelling in Section 3. The right panel of Figure 1 shows that the individuals in the lowest income group have much higher level of disability than the rest of the population. In fact, the average total scores for the five income groups (smallest to largest) are 27.33, 24.41, 23.45, 23.49 and 23.00 respectively. The corresponding standard deviations are 13.14, 10.75, 10.48, 10.93, and 10.17 respectively. This shows a slightly higher variability in the scores for the smallest income group but very similar variability for the remaining groups.

3 Bayesian model descriptions

Let Y_{ij} denote the response of the ith individual on the jth item, where $i = 1, \ldots, n$ and $j = 1, \ldots, J$. The response for the jth item is categorised as one of the K_j possibilities, although in our data example $K_j = K = 5$ for all $j = 1, \ldots, 17$. Thus, each Y_{ij} can take the value y_{ij} where $y_{ij} = 1, 2, \ldots, K_j$. Let $\mathbf{y}$ denote all the observed data y_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Any individual level covariate information, i.e., demographic and socio-economic, will be captured in a m dimensional vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$. In this article we have $m = 6$ corresponding to the continuous covariate age and the dummies for female, and the four upper income quintiles. The effect for lowest income quinitle is set at zero to facilitate comparison with the other income groups.

The GPCM (Muraki, 1992; Li and Baser, 2012) without including any covariate effects is given by:

$$\mathbb{P}\mathrm{r}\left(Y_{ij} = y_{ij} | \theta_i, \alpha_j, \boldsymbol{\beta}_j\right) = \frac{\exp\left\{\sum_{h=1}^{y_{ij}} \alpha_j(\theta_i - \beta_{jh})\right\}}{\sum_{k=1}^{K_j} \exp\left\{\sum_{h=1}^{k} \alpha_j(\theta_i - \beta_{jh})\right\}}, y_{ij} = 1, \ldots, K_j, \tag{1}$$

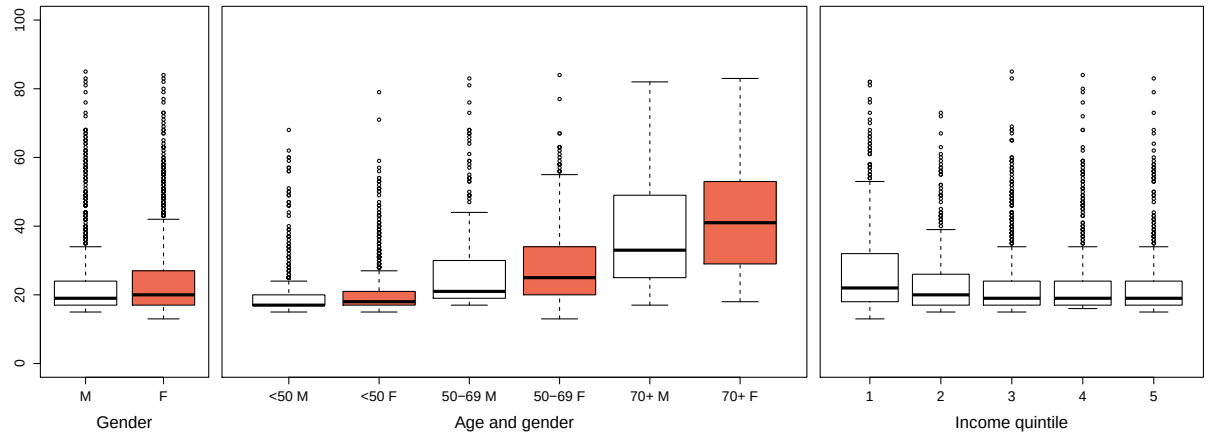

Figure 1: Boxplot of the total scores. Left panel: by gender; middle panel by gender and age groups and right panel by income groups.

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK_j})$. Here, θ_i denotes the latent trait score of the ith individual, α_j denotes the discriminatory power of the jth item and $\boldsymbol{\beta}_j$ are the item specific difficulty parameters, as is common in these type of models often assumed in IRT. The underlying distribution of each Y_{ij} is the multinomial distribution with parameters 1 and the probabilities $\mathbb{P}\mathrm{r}\left(Y_{ij} = y_{ij}|\theta_i, \alpha_j, \boldsymbol{\beta}_j\right)$ as given in (1). Clearly each $\mathbb{P}\mathrm{r}\left(Y_{ij} = y_{ij}|\theta_i, \alpha_j, \boldsymbol{\beta}_j\right) \geq 0$ and $\sum_{y_{ij}=1}^{K_j} \mathbb{P}\mathrm{r}\left(Y_{ij} = y_{ij}|\theta_i, \alpha_j, \boldsymbol{\beta}_j\right) = 1$ for each i and j, as required by the multinomial distribution.

When comparing discrete categories it is customary to nominate a base or reference category. In our modelling the category 1 for each item (j) is taken as the reference category. The model (1) simplifies to the familiar binary logistic regression model when $K_j = 2$ for all $j = 1, \ldots, J$. In this case y_{ij} can take only two values, 1 and 2 and (1) is written as:

$$\begin{aligned}
\mathbb{P}\mathrm{r}\left(Y_{ij} = 1|\theta_i, \alpha_j, \boldsymbol{\beta}_j\right) &= \frac{\exp\left\{\sum_{h=1}^{1} \alpha_j(\theta_i - \beta_{jh})\right\}}{\sum_{k=1}^{2} \exp\left\{\sum_{h=1}^{k} \alpha_j(\theta_i - \beta_{jh})\right\}} \\
&= \frac{\exp\{\alpha_j(\theta_i - \beta_{j1})\}}{\exp\left\{\sum_{h=1}^{1} \alpha_j(\theta_i - \beta_{jh})\right\} + \exp\left\{\sum_{h=1}^{2} \alpha_j(\theta_i - \beta_{jh})\right\}} \\
&= \frac{\exp\{\alpha_j(\theta_i - \beta_{j1})\}}{\exp\{\alpha_j(\theta_i - \beta_{j1})\} + \exp\{\alpha_j(\theta_i - \beta_{j1}) + \alpha_j(\theta_i - \beta_{j2})\}} = \frac{1}{1 + \exp\{\alpha_j(\theta_i - \beta_{j2})\}}
\end{aligned}$$

and $\mathbb{P}\mathrm{r}\left(Y_{ij} = 2|\theta_i, \alpha_j, \boldsymbol{\beta}_j\right) = 1 - \mathbb{P}\mathrm{r}\left(Y_{ij} = 1|\theta_i, \alpha_j, \boldsymbol{\beta}_j\right)$. Thus β_{j1} disappears from the probabilities, as it does for the general model (1). Therefore, we set $\beta_{j1} = 0$ corresponding to the reference category 1 for each item $j = 1, \ldots, J$.

We introduce the covariate information in the above GPCM as follows:

$$\mathbb{P}\mathrm{r}\left(Y_{ij} = y_{ij}|\theta_i, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right) = \frac{1}{G_{ij}} \exp\left\{\sum_{h=1}^{y_{ij}} \alpha_j(\theta_i - \beta_{jh}) + z_{ij}\, \mathbf{x}_i^T \boldsymbol{\gamma}\right\}, \tag{2}$$

where $z_{ij} = I(y_{ij} > 1)$ is the indicator variable taking the value 1 if $y_{ij} > 1$ and 0 otherwise,

and the normalising constant

$$G_{ij} = \sum_{k=1}^{K_j} \exp\left\{\sum_{h=1}^{k} \alpha_j(\theta_i - \beta_{jh}) + I(k > 1)\mathbf{x}_i^T \boldsymbol{\gamma}\right\},$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, J$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$.

Note that we must exclude the regression term $\mathbf{x}_i^T \boldsymbol{\gamma}$ from the reference category ($j = 1$) because if we have $\mathbb{P}\mathrm{r}\left(Y_{ij} = 1 | \theta_i, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right) = \exp\left\{\alpha_j(\theta_i - \beta_{j1}) + \mathbf{x}_i^T \boldsymbol{\gamma}\right\} / G_{ij}$ then the term $\mathbf{x}_i^T \boldsymbol{\gamma}$ is cancelled out in the ratio for every $\mathbb{P}\mathrm{r}\left(Y_{ij} = y_{ij} | \theta_i, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right)$, $y_{ij} = 1, \ldots, K_j$, and the extended model (2) collapses to the original model (1), rendering no likelihood contribution for the regression parameter $\boldsymbol{\gamma}$.

An alternative way to include the covariate information, as has been done by Furr et al. (2016), would be to write the regression component $\mathbf{x}_i^T \boldsymbol{\gamma}$, inside the exponent in (2) as $\sum_{h=1}^{y_{ij}} \alpha_j(\theta_i - \beta_{jh} + z_{ij}\, \mathbf{x}_i^T \boldsymbol{\gamma})$ instead of the proposed $\sum_{h=1}^{y_{ij}} \alpha_j(\theta_i - \beta_{jh}) + z_{ij}\, \mathbf{x}_i^T \boldsymbol{\gamma}$. This alternative formulation allows for multiplicative interaction effect between the item discriminatory parameters α_j and the covariates $\mathbf{x}_i$. However, it causes further identifiability issues in parameter estimation due to the presence of the terms like $\alpha_j \gamma$. The product term, $\alpha_j \gamma$, adds to non-identifiability of the parameters α_j and γ since the product remains unchanged if an arbitrary non-zero constant is multiplied to each α_j and then each γ is divided by the same constant. Below we comment on the identifiability issue further and we also compare these two alternative model specifications in Section 5.

In this paper we do not consider an essentially deterministic model $\theta_i = \mathbf{x}_i^T \boldsymbol{\gamma}$, for the latent trait θ_i as suggested by some authors, see, e.g. de Boeck and Wilson (2004). Instead, we assume normal prior distribution for the latent trait θ_i (see Section 3.1 below) and assume that the covariates provide additional explanatory information in the model.

We now write down the likelihood function needed for our Bayesian model fitting. Let $\boldsymbol{\theta}_n = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{\xi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$. The likelihood function of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}_n$ is given by:

$$L(\boldsymbol{\theta}_n, \boldsymbol{\xi};\, \mathbf{y}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \mathbb{P}\mathrm{r}\left(Y_{ij} = y_{ij} | \theta_i, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right). \tag{3}$$

3.1 Prior and posterior distributions

The Bayesian model is completed by assuming prior distribution for all the parameters. We assume that each $\theta_i \sim N(m_\theta, s_\theta^2)$ independently where m_θ and s_θ^2 are hyper-parameters specified below. The item discriminatory parameters α_j are given independent normal prior distributions with mean m_α and variance s_α^2 but are forced to be on the positive part of the real line.

The item difficulty parameters $\beta_{jh}, h > 1$ are given independent normal prior distributions with mean m_β and variance s_β^2. Components of the regression parameter $\boldsymbol{\gamma}$ are given independent normal prior distributions with mean m_γ and variance s_γ^2. The hyper-parameters of these prior distributions are specified below.

The model is over-parametrised and this poses an identifiability problem since we can multiply each α_j and divide θ_i, β_{jh} by the same constant without changing the likelihood function. To tackle this problem we let $\theta_i \sim N(0, 1)$ as is customary in the literature, see e.g. Albert (1992); Sahu (2002); Sinharay et al. (2006). Typically, we take diffuse prior distributions for regression parameters and so with s_γ^2 large, 10^4 say.

The identifiability of the item discriminatory, $\boldsymbol{\alpha}$ and difficulty, $\boldsymbol{\beta}$ parameters, is also weak, see e.g. Sahu (2002) and Sinharay et al. (2006) which suggests that an informative prior distribution is required for these parameters. Following the above authors we assume $m_\alpha = m_\beta = 0$ and $s_\alpha^2 = s_\beta^2 = 10$. In Section 5, we also compare the proposed model with the model restricting all the discriminatory parameters α's at 1 using the adopted Bayesian model selection criteria.

In our investigation we find that the model parametrised by $\sum_{h=1}^{y_{ij}} \alpha_j(\theta_i - \beta_{jh} + z_{ij}\,\mathbf{x}_i^T\boldsymbol{\gamma})$, labelled as model M4, leads to slower mixing of the MCMC algorithms. Moreover, the model comparison results presented in Table 1 in Section 5 we see that this weakly identified model is not selected by the adopted Bayesian model selection criteria.

The posterior distribution of $\boldsymbol{\theta}_n, \boldsymbol{\xi}$ is now obtained as

$$\pi(\boldsymbol{\theta}_n, \boldsymbol{\xi}|\mathbf{y}) \propto L(\boldsymbol{\theta}_n, \boldsymbol{\xi};\ \mathbf{y})\pi(\boldsymbol{\theta}_n, \boldsymbol{\xi}), \tag{4}$$

where $\pi(\boldsymbol{\theta}_n, \boldsymbol{\xi})$ denotes the joint prior distribution of $\boldsymbol{\theta}_n$ and $\boldsymbol{\xi}$. Hence there are $n + J + \sum_{k=2}^{J} K_j + m$ parameters to estimate.

4 MCMC based inference methods

Model fitting has been performed by using the general purpose software package STAN. The code for implementation and the data are available from the authors upon request and will be published alongside the paper. The adopted STAN software package also facilitates model comparison using the Watanabe information criteria (WAIC) proposed by Watanabe (2010). Gelman et al. (2014) provide a very thorough comparison between the well-known Bayesian model choice criteria including the DIC proposed by Spiegelhalter et al. (2002). Like all information criteria, WAIC is made up of two components, one of which is the effective number of parameters, denoted by p_waic, measuring the complexity of the adopted hierarchical model. The second component assesses the quality of the model-fit and as a result a model with a smaller value of the WAIC is preferred.

Bayesian model checking proceeds by calculating the posterior predictive distribution of a new independent replicate data set $Y_{ij}^{(\text{rep})}$, $i = 1, \ldots, n$, $j = 1, \ldots, J$. The posterior predictive distribution is given by:

$$\mathbb{P}\text{r}(Y_{ij}^{(\text{rep})} = k) = \int \mathbb{P}\text{r}(Y_{ij}^{(\text{rep})} = k|\theta_i, \boldsymbol{\xi})\pi(\theta_i, \boldsymbol{\xi}|\mathbf{y})d\boldsymbol{\theta}d\boldsymbol{\xi} \tag{5}$$

The difficulty in evaluating (5) is easily solved by MCMC methods as suggested by Sinharay et al. (2006), see also Sinharay (2005). First, one obtains a large sample $\boldsymbol{\theta}_n^{(\ell)}, \boldsymbol{\xi}^{(\ell)}$, $\ell = 1, \ldots, L$ for a large value of L, from the posterior distribution $\pi(\boldsymbol{\xi}|\mathbf{y})$. Given each simulated value of $\boldsymbol{\theta}_n^{(\ell)}, \boldsymbol{\xi}^{(\ell)}$, a replicated data set $\mathbf{Y}^{(\text{rep},\,\ell)}$ is generated from the top level multinomial model and for a statistic subsequently $S(\mathbf{Y}^{(\text{rep},\,\ell)})$ is calculated, where $S(\mathbf{y})$ is a suitable summary statistic of the data $\mathbf{y}$. Graphical plots of the replicated $S(\mathbf{Y}^{(\text{rep},\,\ell)})$ with the superimposed value of the observed $S(\mathbf{y})$ are used as informal model checks. For example, here we consider the total number of respondents for each of the 5 catgories of each of the 17 items:

$$S_{jk}(\mathbf{y}) = \sum_{i=1}^{n} I\left(y_{ij} = k\right),\ \ j = 1, \ldots, 17,\ \ k = 1, \ldots, 5, \tag{6}$$

where $I(A)$ denotes the indicator function of its argument A.

More formal model checking is afforded by calculating the posterior predictive 'p-values',

$$\mathbb{P}\text{r}\left(S_{jk}(\mathbf{y}^{(\text{rep})} \geq S_{jk}(\mathbf{y})|\mathbf{y}\right),$$

corresponding to the observed totals $S_{jk}(\mathbf{y})$, $j = 1, \ldots, 17$, $k = 1, \ldots, 5$, see e.g. Sinharay (2005). Here the probabilities are calculated under the discrete posterior predictive distribution defined in (5). Ideally, for a well fitted model these p-values should be close to 0.5 so that the observed totals, $S_{jk}(\mathbf{y})$, $j = 1, \ldots, 17$, $k = 1, \ldots, 5$, are neither under or over predicted by the fitted model.

4.1 Item Information for item ordering

The Fisher's Information for $\theta_i, i = 1, \ldots, n$ from the jth item ($j = 1, \ldots, J$) is given by,

$$I_{ij}(\theta_i, \boldsymbol{\xi}) = -E\left[\frac{\partial^2}{\partial \theta_i^2} \log L(\boldsymbol{\theta}_n, \boldsymbol{\xi}; \mathbf{y})\right],$$

where the expectation is with respect to the distribution of the data $\mathbf{y}$. For notational simplicity we write $k = y_{ij}$ and hence $p_{ijk} \equiv \mathbb{P}\text{r}\left(Y_{ij} = y_{ij} | \theta_i, \alpha_j, \boldsymbol{\gamma}, \boldsymbol{\beta}_j\right)$ in the steps below. By noting the product form of the likelihood $L(\boldsymbol{\theta}_n, \boldsymbol{\xi};\ \mathbf{y})$ in (3) we write,

$$I_{ij}(\theta_i, \boldsymbol{\xi}) = \sum_{k=1}^{K_j} \frac{1}{p_{ijk}} \left[\frac{\partial}{\partial \theta_i} p_{ijk}\right]^2.$$

Note that $I_{ij}(\theta_i, \boldsymbol{\xi})$ depends on the unknown parameters $\boldsymbol{\xi}$. The Bayesian inference paradigm suggests that the posterior distribution of $\boldsymbol{\xi}$ given $\mathbf{y}$, $\pi(\boldsymbol{\xi}|\mathbf{y})$, provides the best information about $\boldsymbol{\xi}$. Hence, the expected value of $I_{ij}(\theta_i, \boldsymbol{\xi})$ with respect to $\pi(\boldsymbol{\xi}|\mathbf{y})$, is the most natural information measure to consider for the jth item provided by the ith individual. Formally, we define,

$$I_{ij}(\theta_i) \equiv E\left(I_{ij}(\theta_i, \boldsymbol{\xi})|\mathbf{y}\right) = \int I_{ij}(\theta_i, \boldsymbol{\xi}) \pi(\boldsymbol{\xi}|\mathbf{y}) d\boldsymbol{\xi}. \tag{7}$$

We use MCMC samples $\boldsymbol{\xi}^{(\ell)}$ for $\ell = 1, \ldots, L$ to estimate $I_{ij}(\theta_i)$ as follows:

$$\hat{I}_{ij}(\theta_i) \equiv \frac{1}{L} \sum_{\ell=1}^{L} I_{ij}(\theta_i, \boldsymbol{\xi}^{(\ell)}). \tag{8}$$

These estimates can be graphically examined for a range of values of θ_i as has been done in the supplementary Figures 8 and 9. Our proposal differs fundamentally from the literature, e.g. Li and Baser (2012) so far. They also obtain $I_{ij}(\theta_i, \boldsymbol{\xi})$ which depends on $\boldsymbol{\xi}$. However, instead of integrating over the uncertainties in the parameter estimates for $\boldsymbol{\xi}$ as done in (7) they replace the unknown $\boldsymbol{\xi}$ by its posterior mean, thus effectively ignoring the uncertainty.

The unknown θ_i in $\hat{I}_{ij}(\theta_i)$ can be integrated over using the posterior samples to obtain

$$\hat{I}_{ij} \equiv \frac{1}{L} \sum_{\ell=1}^{L} I_{ij}(\theta_i^{(\ell)}, \boldsymbol{\xi}^{(\ell)}),$$

which is interpreted as the posterior information for the jth item provided by the ith individual. For item ranking purposes we obtain the item specific Fisher information, FIC, by totaling the information provided by each individual. That is, we obtain

$$I_j = \sum_{i=1}^{n} \int I_{ij}(\theta_i, \boldsymbol{\xi}) \pi(\boldsymbol{\theta}_n, \boldsymbol{\xi}|\mathbf{y}) d\boldsymbol{\theta}_n d\boldsymbol{\xi},$$

which is estimated by $\hat{I}_j = \sum_{i=1}^{n} \hat{I}_{ij}$. Values of the estimated FIC, $\hat{I}_j$, provide the relative information content of the jth item and hence items are ordered according to these.

4.2 Predicting the latent trait score for a new individual

The Bayesian hierarchical GPCM enables prediction of the latent score for a new individual, $n+1$, for whom we may or may not have observed $\mathbf{y}_{n+1} = (y_{n+1,1}, \ldots, y_{n+1,J})$. If $\mathbf{y}_{n+1}$ has not been observed then we treat it as missing data and routine Bayesian methods are available to estimate the unobserved data by sampling from the full conditional distribution of $\mathbf{y}_{n+1}$ given all the parameters and $\mathbf{y}$. Here the parameter vector will include the new latent trait parameter θ_{n+1} and that must be sampled as well conditional on $\mathbf{y}$, the sampled $\mathbf{y}_{n+1}$ and all the parameters. The covariate values for the new individual, $\mathbf{x}_{n+1}$, must be available.

Consider the more interesting case when the observations $\mathbf{y}_{n+1}$ and the covariate values, $\mathbf{x}_{n+1}$ for the new individual are observed after model fitting with data $\mathbf{y}$ from the first n individuals. In this case, it seems to be trivial to estimate (or predict) θ_{n+1} since we can simply augment $\mathbf{y}_{n+1}$ to the nJ dimensional data vector $\mathbf{y}$ to obtain the $(n+1)J$ dimensional data vector and re-fit the model with the additional parameter θ_{n+1}, for which we assume an independent $N(0,1)$ prior distribution. Conceptually, this is very simple but operationally it poses a huge problem when our aim is to estimate the latent scores for a large number of new individuals whose data are observed on a later date after model fitting has been done. The problem arises due to the necessity of re-fitting the model every time a new individual's data becomes available. Below we develop an approximation scheme, which does not require re-fitting, to predict θ_{n+1} using new data $\mathbf{y}_{n+1}$ and $\mathbf{x}_{n+1}$.

We assume that the model has been fitted to the original nJ dimensional data $\mathbf{y}$ and so MCMC iterates $\boldsymbol{\xi}^{(\ell)}, \ell = 1, \ldots, L$ are available from the posterior distribution $\pi(\boldsymbol{\xi}|\mathbf{y})$ in (4). We wish to sample from the marginal distribution of θ_{n+1} given all of the data $\mathbf{y}$ and $\mathbf{y}_{n+1}$. To do so requires evaluating the integral

$$\pi(\theta_{n+1}|\mathbf{y}, \mathbf{y}_{n+1}) = \int \pi(\theta_{n+1}|\boldsymbol{\xi}, \mathbf{y}, \mathbf{y}_{n+1}) \pi(\boldsymbol{\xi}|\mathbf{y}, \mathbf{y}_{n+1}) d\boldsymbol{\xi}. \tag{9}$$

This integral can be evaluated by compositional sampling by first drawing samples $\boldsymbol{\xi}^{*(\ell)} \sim \pi(\boldsymbol{\xi}|\mathbf{y}, \mathbf{y}_{n+1})$ and then drawing $\theta_{n+1}^{(\ell)} \sim \pi(\theta_{n+1}|\boldsymbol{\xi}^{*(\ell)}, \mathbf{y}, \mathbf{y}_{n+1})$ for $\ell = 1, \ldots, L$. However, drawing $\boldsymbol{\xi}^{*(\ell)}$ from $\pi(\boldsymbol{\xi}|\mathbf{y}, \mathbf{y}_{n+1})$ is the re-fitting we aim to avoid. Hence, instead of drawing $\theta_{n+1}^{(\ell)}$ from $\pi(\theta_{n+1}|\boldsymbol{\xi}^{*(\ell)}, \mathbf{y}, \mathbf{y}_{n+1})$ we propose to draw from $\pi(\theta_{n+1}|\boldsymbol{\xi}^{(\ell)}, \mathbf{y}, \mathbf{y}_{n+1})$, avoiding the re-fitting.

We now provide the details for sampling from $\pi(\theta_{n+1}|\boldsymbol{\xi}^{(\ell)}, \mathbf{y}, \mathbf{y}_{n+1})$. The full conditional distribution of θ_{n+1} given $\boldsymbol{\xi}$, $\mathbf{y}$ and $\mathbf{y}_{n+1}$ is given by:

$$\prod_{j=1}^{J} \mathbb{P}\mathrm{r}\left(Y_{n+1,j} = y_{n+1,j}|\theta_{n+1}, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right) \pi(\theta_{n+1}), \tag{10}$$

where $\pi(\theta_{n+1})$ is the $N(0,1)$ prior distribution for θ_{n+1} and following model (2) we have

$$\mathbb{P}\mathrm{r}\left(Y_{n+1,j} = y_{n+1,j}|\theta_{n+1}, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right) = \frac{1}{G_{n+1,j}} \exp\left\{u_{ij}\right\},$$

where

$$u_{ij} = \sum_{h=1}^{y_{n+1,j}} \alpha_j(\theta_{n+1} - \beta_{jh}) + z_{n+1,j}\, \mathbf{x}_{n+1}^T \boldsymbol{\gamma}.$$

In the above, $z_{n+1,j}$ and $G_{n+1,j}$ are similarly defined as in (2), see also the complete conditional distribution for θ_i provided in the Appendix B.

We use a random-walk Metropolis step (see e.g. Tierney (1994)) to sample $\theta_{n+1}^{(\ell)}$ from (10). A proposal sample is drawn from the normal distribution centred at the current value and a variance tuned to have the optimal acceptance rate 0.44 as suggested by Gelman et al. (1996). This proposal is accepted or rejected by evaluating the ratio of the target densities (10) at the current and proposed values. The sampled $\theta_{n+1}^{(\ell)}, \ell = 1, \ldots, L$, values are summarised to provide predictions and their uncertainties for the unobserved latent score θ_{n+1} given the full data $\mathbf{y}, \mathbf{y}_{n+1}$ and the covariate values. The proposed method of sampling $\theta_{n+1}^{(\ell)}$ given the incorrect samples $\boldsymbol{\xi}^{(\ell)}$, instead of the correct $\boldsymbol{\xi}^{*(\ell)}$, effectively constitutes an approximation for the correct marginal posterior distribution (9) of θ_{n+1} given the full data. In our numerical examples we conduct a simulation study to verify this approximation by comparing it to the estimates obtained by re-fitting using the full data set.

5 Results

5.1 Model choice and parameter estimates

The model was run for 1000 iterations after discarding the first 1000 initial iterations at which point MCMC convergence was assessed using the Gelman and Rubin scale reduction factor (Gelman and Rubin, 1992). The WAIC values for four plausible models are presented in Table 1. The descriptive analysis in Section 2 indicates that the main effects of age, income and gender might be significant but not their interactions. Indeed, the main effects model, M2, in Table 1 has a smaller WAIC value than the model without any covariates, M1. This table also contains the WAIC values for the other two models, M3 and M4, discussed in Section 3. Clearly, the main effects model M2 is the best according to the WAIC and henceforth will be the selected model for our purposes in the rest of the paper.

Table 1: WAIC for different models. M1: GPCM without covariates. M2: GPCM with covariates, age, gender, and income as quintiles. M3: GPCM with covariates but all α's set to 1. M4: GPCM with the alternative formulation of covariates as described in Section 3.

	M1	M2	M3	M4
p_waic	1888.5	1734.2	1836.7	1850.0
waic	51432.1	50996.7	54919.7	51291.7

We use the previously described posterior predictive methodology for checking adequacy of the selected model. Corresponding to the observed summaries in Table 5, $S_{jk}(\mathbf{y})$ as defined in (6), we obtain the model based predicted totals at each MCMC iteration and hence the 95% predictive intervals and the posterior predictive p-values. These p-values, reported in Table 2, range from 0.48 to 0.53 and, as expected, are very close to 0.5, although with a slight upward bias, which may be due to the discrete nature of the posterior predictive distribution here. As a further check, we plot the predicted and observed totals along with the 95% intervals in Figure 2. The plot shows a very good agreement between the observed and model predicted totals for each category of each item. Hence, we proceed with this model for making inference.

Table 2: Posterior predictive p-values corresponding to the observed totals in Table 5.

Item	Description	Categories				
		1	2	3	4	5
1	seeing	0.486	0.512	0.495	0.499	0.532
2	hearing	0.505	0.499	0.507	0.514	0.522
3	walking or climbing	0.515	0.512	0.497	0.513	0.481
4	remembering or concentrating	0.514	0.495	0.490	0.484	0.519
5	washing all over or dressing	0.531	0.517	0.509	0.482	0.532
6	communicating	0.492	0.507	0.520	0.515	0.521
7	using hands and fingers	0.501	0.493	0.505	0.530	0.505
8	sleeping	0.515	0.496	0.527	0.506	0.509
9	shortness of breath	0.505	0.521	0.491	0.510	0.510
10	doing household tasks	0.511	0.515	0.506	0.519	0.511
11	providing care for others	0.494	0.529	0.508	0.500	0.503
12	joining community activities	0.513	0.507	0.489	0.515	0.515
13	feeling sad, low or depressed?	0.516	0.514	0.483	0.523	0.522
14	feeling worried or nervous	0.505	0.490	0.501	0.528	0.511
15	getting along with close people	0.518	0.485	0.516	0.529	0.514
16	coping with everything	0.533	0.489	0.516	0.510	0.528
17	bodily aches or pain	0.514	0.502	0.496	0.492	0.495

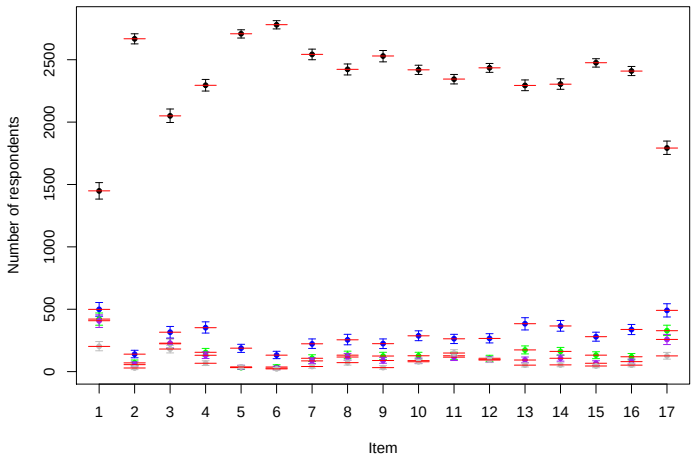

Figure 2: Plot of the predicted totals for the observed $S_{jk}(\mathbf{y}), j = 1, \ldots, 17, k = 1, \ldots, 5$. These observed values are plotted as red lines. Other colours indicate response; black for ones, blue for twos, green for threes, purple for fours, grey for fives. Dots are the expected values with the bars representing 95% predictive intervals.

The estimates for the regression parameters for the selected model are presented in Table 3. The main effect estimate of gender (0.099 for females compared to 0 for males) indicates that the males

Table 3: Parameter estimates for components of $\boldsymbol{\gamma}$ and their 95% credible intervals (CI) using model (2). The first quintile of the 5 level factor income is the baseline, $\gamma_{\text{I2}}, \ldots, \gamma_{\text{I5}}$ are the incremental effects for the upper income quintiles.

Parameter	γ_{Age}	γ_{F}	γ_{I2}	γ_{I3}	γ_{I4}	γ_{I5}
Estimate	0.791	0.099	0.141	–0.079	–0.085	–0.192
95% CI (Lower)	0.738	0.005	–0.018	–0.217	–0.231	–0.333
95% CI (Upper)	0.841	0.193	0.299	0.044	0.059	–0.054

report themselves to be on average healthier than females and, also age is a significant predictor of disability. The parameter estimates for the income effect show that on average people with higher levels of income report lower levels of disability as has been noted in the right panel of Figure 1. However, the difference is only significant between the groups with the highest and lowest levels of income.

In order to assess the effect of including the covariates into the GPCM we provide scatter plots of the parameter estimates obtained from model (2) with the covariates A , G, and I against those obtained using the no-covariate model in Figures 3 and 4. A similar plot for the latent scores $\boldsymbol{\theta}$ is provided in Figure 7 in the online supplement. Without the covariates the GPCM over estimates the $\boldsymbol{\alpha}$ parameters but under estimates the $\boldsymbol{\beta}$ parameters. Such trends, however, are not so discernible in the estimates of the latent scores $\boldsymbol{\theta}$ from the two models. The actual values of the parameter estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are provided in the online supplement. Those reveal that higher estimates for α_j are generally associated with lower values of β_{jh} and vice versa which is expected due to the parameter product $\alpha_j\beta_{jh}$ entering into the GPCM (1). Also note that items 10 and 16 are among the most discriminatory while items 1 and 2 are least discriminatory. This is also confirmed by the item ordering analysis presented below in Section 5.2.

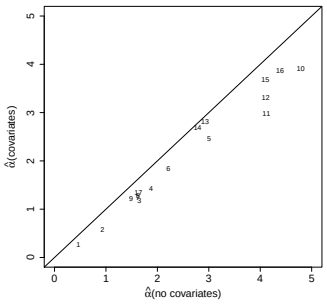

Figure 3: Scatter plot of the estimates of $\boldsymbol{\alpha}$ from the GPCM with and without including the covariates. Item numbers are used as the plotting symbols.

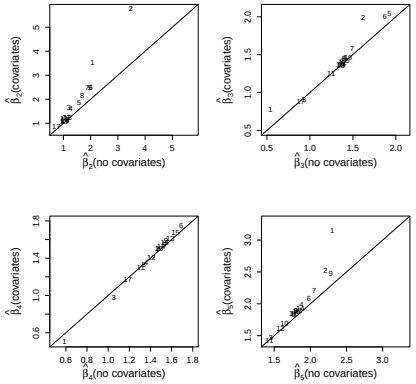

Figure 4: Scatter plot of the estimates of $\boldsymbol{\beta}$ from the GPCM with and without including the covariates. Item numbers are used as the plotting symbols.

5.2 Item ordering

Table 4 provides the ordering of the items according to the FIC. Here items relating to doing household tasks (10), providing care (11), coping with everything (16) joining community (12) rank much higher than the items such as seeing (1) and hearing (2). However, what is surprising is that the psychological item numbers 15 (getting along), 13 (feeling sad) and 14 (feeling worried), all related to mental health, rank much higher than the items 17, 3, and 5 all related to physical health. This is an important finding of this paper.

Table 4: Item ordering according to the estimated FIC values, $\hat{I}_j, j = 1, \ldots, J$ in Column 3. Incremental percentage of the total $\left(\sum_{j=1}^{J} \hat{I}_j\right)$ and the cumulative percentages are shown.

Step	Item added	FIC	% of FIC	Cumulative %
1	10 (household tasks)	2.85	12.71	12.71
2	11 (providing care)	2.57	11.50	24.21
3	16 (coping with everything)	2.52	11.27	35.48
4	12 (joining community)	2.44	10.90	46.37
5	15 (getting along)	2.22	9.93	56.30
6	13 (feeling sad)	1.85	8.27	64.57
7	14 (feeling worried)	1.79	7.98	72.55
8	17 (bodily aches)	1.06	4.73	77.28
9	3 (walking)	0.95	4.22	81.50
10	5 (washing or dressing)	0.91	4.08	85.58
11	4 (remembering)	0.84	3.74	89.32
12	8 (sleeping)	0.7	3.11	92.44
13	7 (using hands)	0.51	2.29	94.73
14	9 (shortness of breath)	0.49	2.17	96.90
15	6 (communicating)	0.47	2.11	99.01
16	2 (hearing)	0.11	0.51	99.52
17	1 (seeing)	0.11	0.48	100

5.3 Prediction of the latent scores for new respondents

In this subsection we illustrate the predictions obtained using the methodology detailed in Section 4.2 using two simulation experiments. In the first experiment we randomly select 50 out of the 3000 individuals and set aside their data for validating the approximation method. We call the selected 50 individuals as 'new' respondents. We fit the model using the data for the remaining 2950 individuals and then predict the latent score of each of the 50 new respondents. In so doing we use the response data y_{ijk} for each of these 50 individuals and their covariate values too. We also obtain the 95% prediction intervals associated with these predictions in each case. We judge the accuracy of the approximate prediction values by comparing them with the estimates obtained by fitting the model to the data from all 3000 respondents. Figure 5 compares the approximate predictions and the associated 95% prediction intervals. The actual predictions are seen to be very close to the approximate ones. This figure also shows that the prediction intervals are tighter for the individuals at the higher end of the latent trait scores, which is also observed by the much larger second experiment described below.

To provide further evidence on the accuracy of the approximation method, the simulation experiment repeats the first one by setting aside data for 1000 randomly selected individuals and fitting the model with data from the remaining 2000 individuals. Figure 6 shows that the approximate scores for the 1000 'new individuals' scatter very tightly around the estimates obtained by fitting the model to the full data set. This figure also reveals that there is better agreement between the approximate and actual values at the higher end of the latent scores than at the lower end, as similarly noted in Figure 5. This is intuitively justified since the individuals with the higher latent scores provide better information and hence their abilities are better estimated with lower levels of uncertainties.

6 Discussion

This paper has set out to achieve three inferential tasks when the GPCMs are employed. The first task enables estimation of the item parameters adjusted for the covariate effects. Inference for the covariate effects has been illustrated with the Model Disability Survey data from the WHO. A Bayesian approach based on a single model, in contrast to a stage-wise procedural estimation method, allows us to accurately assess the uncertainties not only for the item parameters but also for the regression parameters.

The second inferential task has been to rank the items so that a brief version of the MDS with fewer items can be prepared. Using an expected FIC we have developed a method for ranking the items. It is up to the practitioner to decide how many items can be afforded due to cost considerations in the reduced survey and we acknowledge that there may be other practical considerations which may influence the final item choice. The proposed method will guide item selection based on a desired percentage of information that must be present in the reduced survey.

The third inferential task is the MCMC based methodology to predict the latent trait scores of new individuals whose data are observed after model fitting has already been performed. The methodology uses all the relevant covariate information of each new individual so that the best possible Bayesian estimates are obtained. The proposed prediction methodology has been empirically verified by re-estimating the latent trait scores of a large number (1000) of new individuals by fitting the model to the full data set. Close agreement between the predicted scores and the estimated scores based on all the data shows the effectiveness of the new methodology.

These three methodological extensions allowed us to extract a lot more information from the data than what has been possible before, e.g. Sabariego et al. (2015). Using a unified model it has been concluded that the main effects of gender, age and income are all significant in the presence of, hence accounting for, the latent latent trait (θ), item discriminatory (α) and item difficulty (β) parameters. In addition, the main advantage of the unified model also lies in its ability to make coherent inference on item ordering and latent score prediction for new individuals. By eliminating a stage-wise approach for the three different inferential tasks, the developed Bayesian methodology proposes a rigorous and coherent inference framework wherever GPCM models are to be used in practice. This framework ensures coherency by having the correct and mutually consistent levels of uncertainty in the three different inferential tasks.

The methodological developments, though illustrated using the GPCM, can also be used with the simpler one, two and three parameter item response models. Implementation of these models using general purpose Bayesian software packages such as, Stan Development Team (2015); Thomas et al. (2006), is relatively simple and here we have provided STAN code used in this paper.

This paper has not considered verification of the three most important assumptions inherent in IRT, namely uni-dimensionality, local independence and monotonicity. Checking these assumptions for the

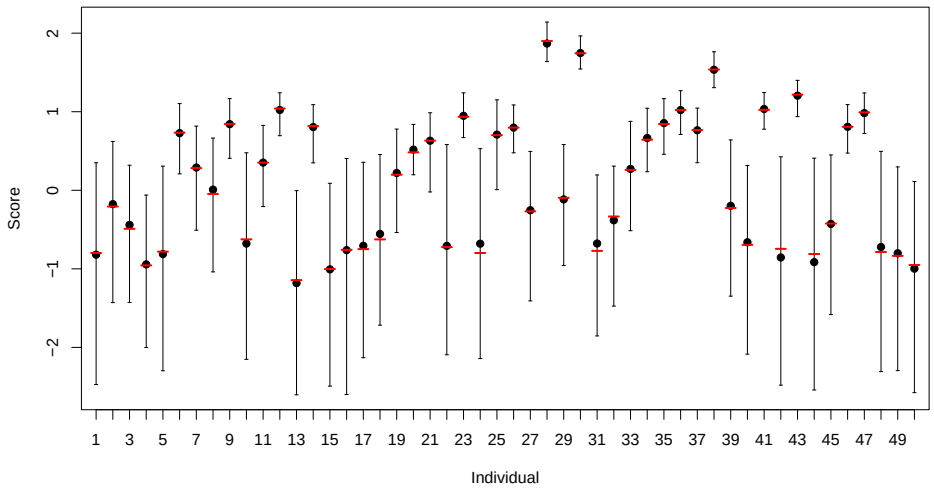

Figure 5: Plot of the predicted θ_i (black dots) for 50 presumed new respondents using the proposed approximation method in Section 4.2. The line segments represent the associated 95% predictive intervals and red lines indicate the estimates obtained using the full data set including the new respondent.

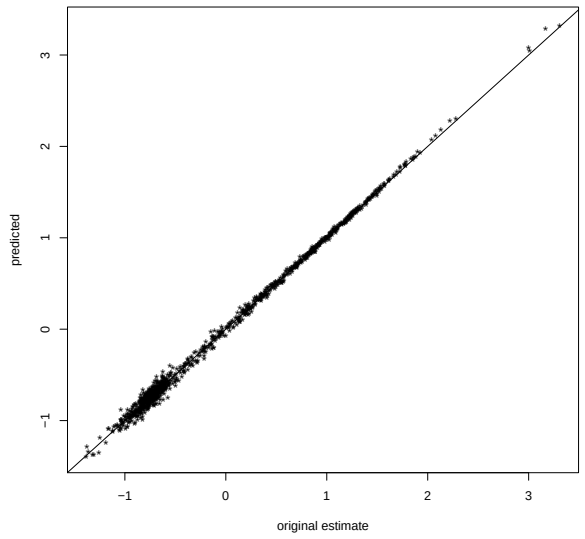

Figure 6: Scatter plot of the predicted θ_i for a new set of 1000 presumed new respondents using the proposed approximation method in Section 4.2 against the original estimates obtained using the full data set including the new respondent. The line $y = x$ is superimposed.

MDS data used here has been discussed in Sabariego et al. (2015) using non-Bayesian tests of hypotheses such as bi-factor analysis and informal graphical methods. Bayesian solutions to these problems are based on posterior predictive checks discussed in Section 4 after model fitting has been completed. See Sinharay (2005) and Sinharay et al. (2006) for further details regarding the posterior predictive checks.

This paper, as the title suggests, has considered the GPCM only, but not its competitors such as the graded response model (GRM), (Samejima, 1969). Recently, Silva et al. (2019) have illustrated some small differences, as measured by several Bayesian model choice criteria, between the GPCMs and GRMs using simulation studies and a real data example. It will be worthwhile to make such comparisons for the current data set in a future article. In addition, such future research efforts can also consider modelling the analysed data set in a multi-group framework.

The developed methods can be applied in medical fields where comparative quantification of health status is required to compare persons at a cross section and over time. In clinical medicine and health surveys, it is imperative to quantify the level of health of a given individual that can be aggregated to population levels. This is important to measure the impact of interventions both at an individual as well as the population level. It is also necessary to monitor changes over time. This metric to quantify health status needs therefore to be comparable across population and over time. Furthermore, since all the observations may not be available at the same time, an analytical strategy that allows one to scale these measurements at different points in time or different populations, on the same scale is crucial. Health states are often measured as an individual's execution of a task or action in a range of domains that is then aggregated into a composite vector of health status, see e.g. Salomon et al. (2003). Our approach to the analysis of data from a national population survey demonstrates the feasibility of quantifying the levels of health status and addresses the above mentioned challenges.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics 17*(3), 251–269.

Bartolucci, F., S. Bacci, and M. Gnaldi (2014). MultiLCIRT: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis 71*, 971 – 985.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*, pp. 397–472. Addison-Wesley, Reading (Mass).

Bock, R. D. and M. Aitkin (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika 46*, 443–459.

Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software 36*(1), 1–34.

de Boeck, P. and M. Wilson (2004). *Explanatory Item Response Models*. Springer Verlag: New York.

Falk, C. F. and L. Cai (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika 81*(2), 434–460.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Fox, J.-P. and C. A. W. Glas (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika 66*(2), 271–288.

Furr, D. C., S. Y. Lee, J.-H. Lee, and S. Rabe-Hesketh (2016). Two-parameter logistic item response model. Technical report, University of California at Berkeley.

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

Gelman, A., G. O. Roberts, and W. R. Gilks (1996). Efficient metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. (Eds.), *Bayesian Statistics 5*, pp. 599–607. Oxford University Press.

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

Glas, C. A. W. (2005). Book review: de boeck, p., andwilson, m. eds. explanatory item responsemodels: A generalized linear and nonlinear approach. new york, ny: Springer, 2004. *Journal of Educational Measurement 42*, 303–307.

Glas, C. A. W. and R. R. Meijer (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement 27*(3), 217–233.

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software 20*(10), 1–24.

Johnson, T. R. and K. M. Kuhn (2015). Simulation-based Bayesian inference for latent traits of item response models: Introduction to the ltbayes package for R. *Behavior Research Methods 47*(4), 1309–1327.

Karabatsos, G. (2017). Bayesian nonparametric irt. In W. van der Linden (Ed.), *Handbook Of Item Response Theory: Models, Statistical Tools, and Applications, Volume 1*, pp. Chapter 19. New York: Taylor & Francis.

Li, Y. and R. Baser (2012). Using R and WinBUGS to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine 31*(18), 2010–2026.

Li, Y. H. and R. W. Lissitz (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement 41*(2), 85–117.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika 51*, 177–195.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement 16*(2), 159–176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement 17*, 351–363.

Patz, R. J. and B. W. Junker (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics 24*(4), 342–366.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. (1991). Kernal smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika 56*(4), 611–630.

Rasch, G. (1961). On general laws and the meaning of the measurement in psychology. In *Proceedings of the 4th Berkley Symposium on Mathematical Statistics.*, Volume 4, pp. 321–334. University of California Press.

Rizopoulos, D. (2006). ltm: An r package for latent variable modeling and item response theory analyses. *Journal of Statistical Software 17*(5), 1–25.

Sabariego, C., C. Oberhauser, A. Posarac, J. Bickenbach, N. Kostanjsek, S. Chatterji, A. Officer, M. Coenen, L. Chhan, and C. A. (2015). Measuring disability: Comparing the impact of two data collection approaches on disability rates. *International Journal of Environmental Research and Public Health 12*(9), 10329–51.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation 72*, 217–232.

Salomon, J., C. Mathers, S. Chatterji, R. Sadana, T. Üstün, and C. Murray (2003). Quantifying individual levels of health: definitions, concepts and measurement issues. In M. CJL and E. D (Eds.), *Health Systems Performance Assessment: Debates, Methods and Empricism*, Chapter 26, pp. 301–318. Geneva: World Health Organisation.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Psychometric Monographs, 17, 1100.

Silva, M. A., J. L. Bazan, and A. C. Huggins-Manley (2019). Sensitivity analysis and choosing between alternative polytomous irt models using bayesian model comparison criteria. *Communications in Statistics - Simulation and Computation 48*(2), 601–620.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement 42*(4), 375–394.

Sinharay, S., M. S. Johnson, and H. S. Stern (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement 30*(4), 298–321.

Spiegelhalter, S. D., N. G. Best, B. P. Carlin, and A. V. D. Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B 64*(4), 583–639.

Stan Development Team (2015). *Stan modeling language: Users guide and reference manual*. Columbia, New York: Columbia University.

Swaminathan, H. and J. A. Gifford (1986). Bayesian estimation in the three parameter logistic model. *Psychometrika 51*, 589–601.

Thomas, A., B. O Hara, U. Ligges, and S. Sturtz (2006). Making `bugs` open. `R News` *6*, 12–17.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics 22*(4), 1701–1728.

Tsutakawa, R. K. and H. Y. Lin (1986). Bayesian estimation of item response curves. *Psychometrika 51*, 251–267.

Verhagen, J. and J.-P. Fox (2013). Longitudinal measurement in health related surveys. a bayesian joint growth model for multivariate ordinal responses. *Statistics in Medicine 32*, 2988–3005.

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

Supplementary materials

Appendix A: Contains additional results with larger tables and graphs.

Appendix B: Full Conditional distributions needed in our implementation.

Data: All data files can be found in the folder labelled data_files. A full description is given in the README.txt file contained therein.

R code: The R code used to call STAN and to do the post-processing can be found in the folder labelled R_files. A full description of the files is given in README.txt.

STAN code: The STAN code used to fit the models can be found in the folder labelled stan_files. A full description of the files is given in README.txt.

Appendix A: Additional results with larger tables and graphs.

Table 5: For each of the $J = 17$ items a brief description is provided in Column 2. Items 1-16 started with "How much difficulty do you have..." and item 17 asked, "How much bodily aches or pain do you have?" Columns 3-7 provide the number of respondents selecting each of the $K = 5$ categories out of $n = 3000$ participants. Some participants did not respond to certain items, hence there are some row totals that are less than 3000.

Item	Description	Category					Total
		1	2	3	4	5	
1	seeing	1449	499	421	408	202	2979
2	hearing	2668	140	70	56	29	2963
3	walking or climbing	2050	316	223	228	181	2998
4	remembering or concentrating	2295	352	155	131	67	3000
5	washing all over or dressing	2708	188	32	35	37	3000
6	communicating	2781	132	37	28	22	3000
7	using hands and fingers	2543	223	107	86	41	3000
8	sleeping	2423	255	132	117	72	2999
9	shortness of breath	2530	224	125	89	32	3000
10	doing household tasks	2419	287	127	88	79	3000
11	providing care for others	2345	263	127	116	149	3000
12	joining community activities	2435	266	107	95	94	2997
13	feeling sad, low or depressed?	2294	384	174	93	52	2997
14	feeling worried or nervous	2304	366	162	107	54	2993
15	getting along with close people	2476	280	132	67	45	3000
16	coping with everything	2409	338	119	80	52	2998
17	bodily aches or pain	1792	491	328	258	126	2995

Table 6 provides the parameter estimates and their 95% credible intervals for the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters. All of the α parameters for item discrimination are significant since the 95% credible interval for each α_j does not include the value 1. Similarly all β_{jh} for $h > 1$ are significant since the 95% credible interval

Table 6: Estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and their 95% credible intervals (CI) using model (2) with A , G and I.

j	α_j	β_{j2}	β_{j3}	β_{j4}	β_{j5}
1	0.27(0.23, 0.31)	3.53(2.78, 4.41)	0.78(0.34, 1.24)	0.51(–0.01, 0.98)	3.15(2.47, 3.86)
2	0.59(0.48, 0.69)	5.79(4.89, 6.87)	1.99(1.52, 2.49)	1.5(0.93, 2.06)	2.53(1.75, 3.32)
3	1.19(1.07, 1.31)	1.64(1.43, 1.88)	0.91(0.76, 1.08)	0.98(0.81, 1.16)	1.47(1.28, 1.67)
4	1.43(1.27, 1.59)	1.63(1.45, 1.82)	1.47(1.32, 1.63)	1.35(1.19, 1.53)	1.99(1.78, 2.22)
5	2.46(2.13, 2.83)	1.88(1.73, 2.06)	2.05(1.86, 2.25)	1.55(1.32, 1.74)	1.83(1.6, 2.07)
6	1.84(1.57, 2.12)	2.48(2.23, 2.77)	2(1.78, 2.24)	1.75(1.45, 2.05)	2.08(1.73, 2.44)
7	1.24(1.08, 1.41)	2.48(2.2, 2.8)	1.58(1.36, 1.79)	1.49(1.24, 1.75)	2.21(1.89, 2.54)
8	1.29(1.15, 1.44)	2.15(1.92, 2.4)	1.42(1.21, 1.61)	1.33(1.13, 1.55)	1.89(1.66, 2.14)
9	1.21(1.07, 1.37)	2.51(2.24, 2.8)	1.45(1.26, 1.66)	1.59(1.35, 1.85)	2.47(2.13, 2.84)
10	3.9(3.47, 4.36)	1.14(1.06, 1.23)	1.37(1.29, 1.46)	1.5(1.41, 1.6)	1.69(1.57, 1.8)
11	2.98(2.67, 3.35)	1.19(1.09, 1.29)	1.26(1.16, 1.36)	1.3(1.2, 1.41)	1.42(1.31, 1.53)
12	3.3(2.93, 3.73)	1.24(1.15, 1.35)	1.4(1.31, 1.5)	1.41(1.3, 1.52)	1.62(1.51, 1.73)
13	2.82(2.53, 3.14)	1.07(0.99, 1.17)	1.36(1.27, 1.45)	1.61(1.49, 1.73)	1.9(1.74, 2.06)
14	2.68(2.4, 3)	1.12(1.02, 1.22)	1.37(1.27, 1.47)	1.52(1.41, 1.65)	1.93(1.78, 2.09)
15	3.68(3.28, 4.12)	1.24(1.15, 1.33)	1.42(1.34, 1.51)	1.68(1.56, 1.8)	1.89(1.75, 2.04)
16	3.88(3.44, 4.35)	1.11(1.04, 1.19)	1.46(1.38, 1.55)	1.57(1.47, 1.68)	1.86(1.73, 2)
17	1.34(1.21, 1.46)	0.88(0.73, 1.03)	0.88(0.77, 0.99)	1.17(1.03, 1.31)	1.85(1.67, 2.02)

for each does not include the value 0. Recall that β_{j1} has been set to zero. The parameter estimates also reveal that higher estimates for α_j are generally associated with lower values of β_{jh} and vice versa which is expected due to the parameter product $\alpha_j\beta_{jh}$ entering into the GPCM (1).

The actual values of the FIC are plotted as boxplots in Figure 10. The plot is in broad agreement with the estimates of the item discriminatory parameter estimates reported earlier in Figure 3 where item 10 is the most discriminatory while item 1 is the least discriminatory. The plots show more extreme values for the top 5 high ranking items than the bottom ranking items. The top ranking items are expected to have large values of FIC as their information content is much higher and there may be many individuals for whom these items are very informative. In the lower half of the ranked items we also find item numbers 5 (washing or dressing) and 6 (communicating) having many extreme values. This may be explained by the presence of many elderly individuals for whom washing or dressing is much more problematic. They may have more trouble in communicating as well. The full Bayesian inference extension of this article has enabled us to perform such a detailed level of analysis, which is in contrast to a plug-in Bayesian methodology.

Figure 8 provides density plots of the FIC for items 1 and 2 providing the least information (top row) and items 10 and 11 providing the most information (bottom row) over a range of values for θ for a typical 50 year old woman with median income. This figure further explores the full capability of the MCMC based Bayesian method as it reveals how informative is each item according to each typical individual. Figure 9 provides similar plots for men in different income category. These show the gender effect already seen in the main manuscript.

Appendix B: Full conditional distributions needed in our implementation

In what follows we give details of the full conditional distributions needed to sample from the joint posterior distribution (4). Each distribution is unidimensional although jointly updating two or more

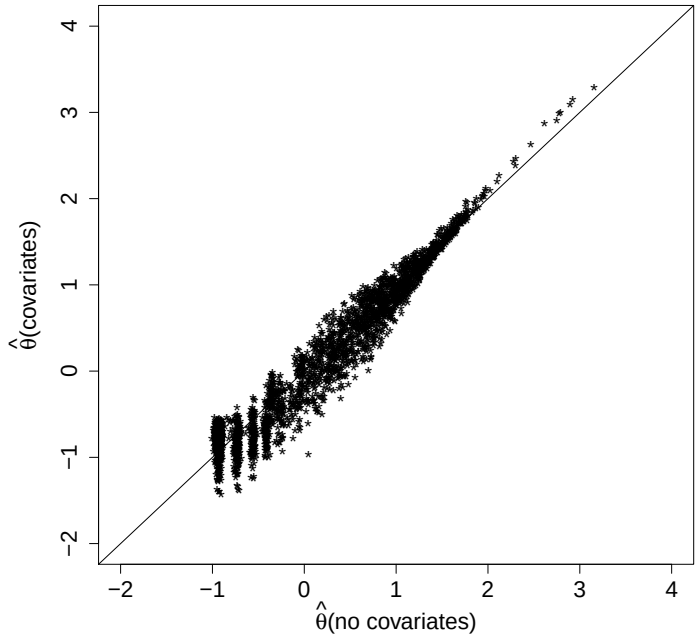

Figure 7: Scatter plot of the estimates from $\boldsymbol{\theta}$ from the GPCM with and without including the covariates.

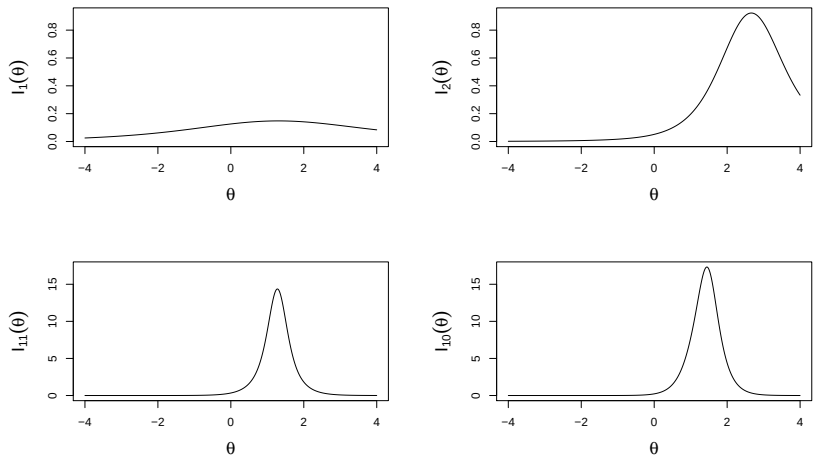

Figure 8: Density plots of the Fisher information $\hat{I}_{ij}(\theta_i)$ defined in (8) for items 1 and 2 providing the least information (top row) and items 11 and 10 providing the most information (bottom row) over a range of values for θ for a 50 year old woman with median income.

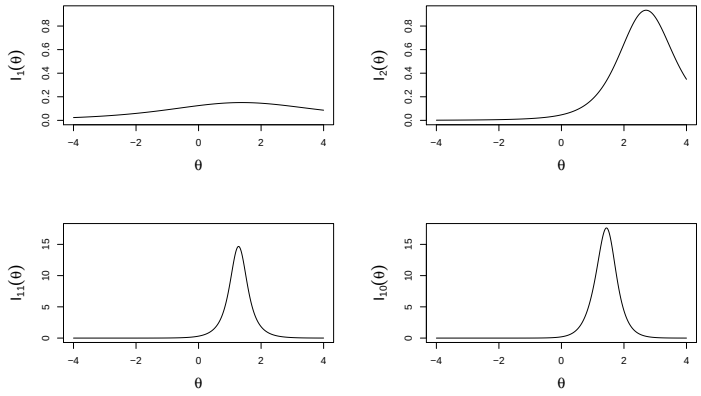

Figure 9: Density plots of the Fisher information for each item over a range of values for θ, found using the plugin estimates for a 50 year old man with median income.

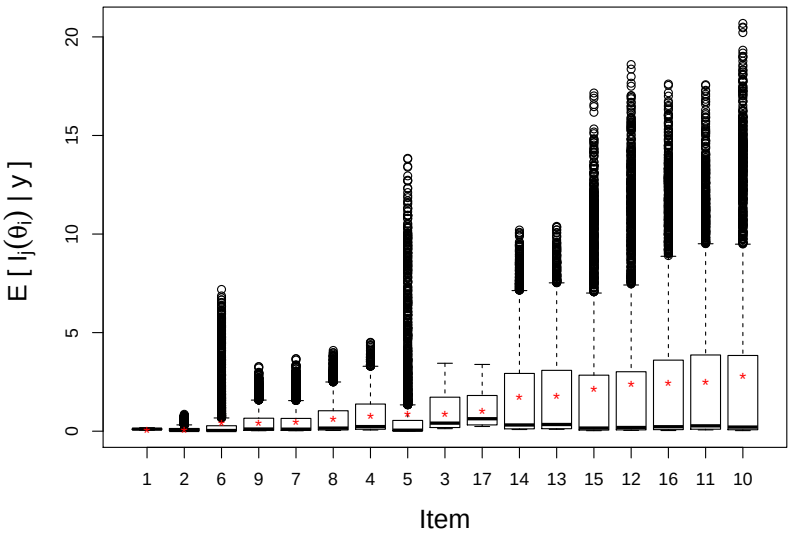

Figure 10: Items are ranked according to the estimated FIC values. The boxplot for each of the 17 items is based on $\hat{I}_{ij}$, $i = 1, \ldots, 3000$. The mean of the 3000 $\hat{I}_{ij}$ is shown as a *.

parameters at once is of course possible. We use the notation $\boldsymbol{\theta}_{-i}$ to mean the vector $\boldsymbol{\theta}$ without θ_i. The vectors $\boldsymbol{\alpha}_{-j}$, $\boldsymbol{\beta}_{-jh}$ and $\boldsymbol{\gamma}_{-l}$ have similar interpretations. Further, we let $\pi(\cdot)$ denote the prior distribution of its argument. The full conditional distribution of θ_i, $i = 1, \ldots, n$, is

$$\begin{aligned}\pi(\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{y}) &\propto \prod_{j=1}^{J} \Pr\left(Y_{ij} = y_{ij}|\theta_i, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right) \pi(\theta_i) \\ &\propto \prod_{j=1}^{J} \frac{1}{G_{ij}} \exp\left\{y_{ij}\alpha_j\theta_i - \tfrac{1}{2}\,\theta_i^2\right\}.\end{aligned}$$

The full conditional distribution of α_j, $j = 1, \ldots, J$ is

$$\begin{aligned}\pi(\alpha_j|\boldsymbol{\theta}, \boldsymbol{\alpha}_{-j}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{y}) &\propto \prod_{i=1}^{n} \Pr\left(Y_{ij} = y_{ij}|\theta_i, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right) \pi(\alpha_j) \\ &\propto \prod_{i=1}^{n} \frac{1}{G_{ij}} \exp\left\{\sum_{h=1}^{y_{ij}} \alpha_j(\theta_i - \beta_{jh}) - \frac{1}{2\,s_\alpha^2}(\alpha_j - m_\alpha)^2\right\},\end{aligned}$$

for $\alpha_j \in [0, \infty)$, zero otherwise. The full conditional distribution of β_{jh}, $j = 1, \ldots, J$, $h = 2, \ldots, K_j$ is

$$\begin{aligned}\pi(\beta_{jh}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{-jh}, \boldsymbol{\gamma}, \mathbf{y}) &\propto \prod_{i=1}^{n} \Pr\left(Y_{ij} = y_{ij}|\theta_i, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}\right) \pi(\beta_{jh}) \\ &\propto \prod_{i=1}^{n} \frac{1}{G_{ij}} \exp\left\{v_{ij}\right\},\end{aligned}$$

where

$$v_{ij} = -\alpha_j I(y_{ij} = h)\beta_{jh} - \frac{1}{2\,s_\beta^2}(\beta_{jh} - m_\beta)^2,$$

and $I(y_{ij} = h)$ takes the value 1 if $y_{ij} = h$ and 0 otherwise.

The full conditional distribution of γ_l, $l = 1, \ldots, m$, is given by

$$\pi(\gamma_l|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}_{-l}, \mathbf{y}) \propto L(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma};\ \mathbf{y}) \exp\left\{-\frac{1}{2\,s_\gamma^2}(\gamma_l - m_\gamma)^2\right\},$$

where $L(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma};\ \mathbf{y})$ is the likelihood function given in (3).