Journal of Statistical Software

November 2008, Volume 28, Issue 10. http://www.jstatsoft.org/

A MATLAB Package for Markov Chain Monte Carlo with a Multi-Unidimensional IRT Model

Yanyan Sheng
Southern Illinois University-Carbondale

Abstract

Unidimensional item response theory (IRT) models are useful when each item is designed to measure some facet of a unified latent trait. In practical applications, items are not necessarily measuring the same underlying trait, and hence the more general multi-unidimensional model should be considered. This paper provides the requisite information and description of software that implements the Gibbs sampler for such models with two item parameters and a normal ogive form. The software developed is written in the MATLAB package **IRTmu2no**. The package is flexible enough to allow a user the choice to simulate binary response data with multiple dimensions, set the number of total or burn-in iterations, specify starting values or prior distributions for model parameters, check convergence of the Markov chain, as well as obtain Bayesian fit statistics. Illustrative examples are provided to demonstrate and validate the use of the software package.

Keywords: multi-unidimensional IRT, two-parameter normal ogive models, MCMC, Gibbs sampling, Gelman-Rubin R, Bayesian DIC, posterior predictive model checks, MATLAB.

1. Introduction

Modeling the interaction between persons and items at the item level for binary or polytomous response data, item response theory (IRT) is a popular approach for solving various measurement problems, and has been found useful in a wide variety of applications in education and psychology (e.g., Embretson and Reise 2000; Kolen and Brennan 1995; Lord 1980; Wainer *et al.* 2000) as well as in other fields (e.g., Bafumi *et al.* 2005; Bezruckzo 2005; Chang and Reeve 2005; Feske *et al.* 2007; Imbens 2000; Sinharay and Stern 2002). For dichotomously scored items, IRT relates the probabilistic 0/1 responses with the person's latent trait(s), $\boldsymbol{\theta}_i$, and the item's characteristics, $\boldsymbol{\xi}_j$, in a way that

$$P(y = 1) = f(\boldsymbol{\theta}_i, \boldsymbol{\xi}_j),$$

where f can be a probit or logit function. It is noted that in the IRT literature, the probit model is generally referred to as the normal ogive model, and the logit model is referred to as the logistic model. Common IRT models assume one θ_i parameter for each person, and are referred to as the unidimensional models, signifying that each test item measures some facet of the unified latent trait. It is necessary that a test assuming one dimension will not be affected by other dimensions. However, psychological processes have constantly been found to be more complex and an increasing number of measurements in education, psychology or other fields assess a person on more than one latent trait, or require response processes with different cognitive components. With regard to this, allowing separate inferences to be made about persons for each distinct latent dimension being measured, multidimensional IRT models have shown promise for dealing with such complexity in situations where multiple traits are required in producing the manifest responses to an item (Reckase 1997). Often, however, a test involves several latent traits and each item measures one of them. The multidimensional model specific for this scenario is referred to as the so-called "multi-unidimensional IRT model" (Sheng and Wikle 2007). In the literature, this model has been called the multidimensional model with simple structure (McDonald 1999) or the between-items multidimensional model (Adams *et al.* 1997). The shorter term "multi-unidimensional" is adopted in this paper to account for the fact that the overall test involves multiple traits, whereas each subtest is unidimensional. The multi-unidimensional model can be viewed as a special case of the multidimensional model. Their difference lies in the consideration whether individual test items measure one or multiple cognitive component(s), which is best understood from the factor analytic perspective. Analogously, multidimensionality is assumed if one believes that each test item has nonzero loadings on all factors extracted, and multi-unidimensionality is considered when the factor solution achieves a simple structure (that is, each item has nonzero loadings on only one of the factors). With respect to the latter, the latent dimensional structure should be specified a priori based on theoretical considerations. In particular, one has to decide on the number of latent traits the test is designed to measure and the specific items involved in measuring each distinct trait.

In situations where a test consists of several subtests with each being unidimensional, if it is a priori clear that all the latent traits being measured are highly correlated, a unidimensional IRT model may be assumed for the overall test because each specific trait can be viewed as some aspect of the unified latent dimension. On the other hand, if it is believed that the traits are not correlated, one may fit a unidimensional model for each subtest, as the subtests can be assumed to be independent from each other. These two approaches using the unidimensional model are, however, restricted in situations where the latent dimensions correlate in other ways. In many applications, prior information on the intertrait relation is not readily available. It is hence difficult to decide whether a unidimensional model for the overall test or for individual subtests is appropriate. The more general multi-unidimensional model, applicable in situations where the latent traits have various levels of correlations, should then be considered. Its advantage is further illustrated in a later section of this paper.

In IRT, parameter estimation offers the basis for its theoretical advantages, and hence has been a major concern in the application of IRT models. As the influence of items and persons on the responses is modeled by distinct sets of parameters, simultaneous estimation of these parameters results in statistical complexities in the estimation task, which have made estimation procedure a primary focus of psychometric research over decades (Birnbaum 1969; Bock and Aitkin 1981; Molenaar 1995). With the enhanced computational technology, recent

attention has been focused on a fully Bayesian approach using Markov chain Monte Carlo (MCMC; e.g., Chib and Greenberg 1995) simulation techniques, which are extremely general and flexible and have proved useful in practically all aspects of Bayesian inferences, such as parameter estimation or model comparisons. One of the simplest MCMC algorithms is Gibbs sampling (Casella and George 1992; Gelfand and Smith 1990; Geman and Geman 1984). The method is straightforward to implement when each full conditional distribution associated with a particular multivariate posterior distribution is a known distribution that is easy to sample. Gibbs sampling has been applied to common unidimensional models (Albert 1992; Johnson and Albert 1999) using the data augmentation idea of Tanner and Wong (1987). Lee (1995) further extended the approach of Albert (1992) and developed the Gibbs sampling procedure for the two-parameter normal ogive (2PNO) multi-unidimensional model, which was found to be more flexible and efficient compared with the conventional unidimensional model (Sheng and Wikle 2007). As a natural extension of the conventional 2PNO IRT model, the 2PNO multi-unidimensional model generalizes the unidimensional model to be on the multi-unidimensional structure. It is hence considered as the standard conceptualization of multi-unidimensionality in IRT.

This paper provides a MATLAB (The MathWorks, Inc. 2007) package that implements Gibbs sampling for the 2PNO multi-unidimensional IRT model with the option of specifying noninformative or informative priors for item parameters. Section 2 reviews the model and briefly describes the MCMC algorithm implemented in the package **IRTmu2no**. In Section 3, a brief illustration is given of Bayesian model choice or checking techniques for testing the adequacy of a model. The package **IRTmu2no** is introduced in Section 4, where a description is given of common input and output variables. In Section 5, illustrative examples are provided to demonstrate the use of the source code. Finally, a few summary remarks are given in Section 6.

It has to be noted that more complicated MCMC procedures have to be adopted for the logistic form of IRT models (e.g., Patz and Junker 1999a,b). As Gibbs sampling is relatively easier to implement, and the logistic and normal ogive forms of the IRT model are essentially indistinguishable in model fit or parameter estimates given proper scaling (Birnbaum 1968; Embretson and Reise 2000), MCMC procedures for logistic models are not considered in this paper.

2. Model and MCMC algorithm

Multi-unidimensional models allow separate inferences to be made about a person for each distinct dimension being measured by a test item while taking into consideration the relationship between all latent traits measured by the overall test. Suppose a K-item test consists of m subtests, each containing k_v dichotomous (0-1) items, where $v = 1, 2, \ldots, m$. Let y_{vij} denote the ith person's response to the jth item in the vth subtest, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k_v$. With a probit link, the 2PNO multi-unidimensional model is defined as

$$P(y_{vij} = 1|\theta_{vi}, \alpha_{vj}, \beta_{vj}) = \Phi(\alpha_{vj}\theta_{vi} - \beta_{vj}) = \int_{-\infty}^{\alpha_{vj}\theta_{vi}-\beta_{vj}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \tag{1}$$

(e.g., Lee 1995; Sheng and Wikle 2007), where θ_{vi} is a scalar person trait parameter in the vth latent dimension, α_{vj} is a positive scalar slope parameter representing the item discrimination,

and β_{vj} is a scalar intercept parameter that is related to the location in the vth dimension where the item provides maximum information.

To implement Gibbs sampling to the model, an augmented continuous variable Z is introduced so that $Z_{vij} \sim N(\eta_{vij}, 1)$ (Albert 1992; Lee 1995; Tanner and Wong 1987), where $\eta_{vij} = \alpha_{vj}\theta_{vi} - \beta_{vj}$. Denote each person's latent traits measured by all test items as $\boldsymbol{\theta}_i = (\theta_{1i}, \ldots, \theta_{mi})'$, and specify a multivariate normal prior distribution for them so that $\boldsymbol{\theta}_i \sim N_m(\mathbf{0}, \mathbf{P})$, where $\mathbf{P}$ is a covariance matrix with the variances being fixed at 1. It is noted that the proper multivariate normal prior for θ_{vi} with their location and scale parameters specified to be 0 and 1, respectively, ensures unique scaling and hence is essential in resolving a particular identification problem for the model (see e.g. Lee 1995, for a description of the problem). Further, it follows that the off-diagonal element of $\mathbf{P}$ is the correlation ρ_{st} between θ_{si} and θ_{ti}, $s \neq t$. One may note that when $\rho_{st} = 1$ for all s, t, the model reduces to the unidimensional 2PNO model, whose probability function is defined as

$$P(y_{ij} = 1|\theta_i, \alpha_j, \beta_j) = \Phi(\alpha_j\theta_i - \beta_j), \qquad i = 1, \ldots, n, \quad j = 1, \ldots, K. \tag{2}$$

In addition, when $\rho_{st} = 0$ for all s, t, the model is actually equivalent to fitting a 2PNO unidimensional model for each subtest. Hence, the two approaches using the unidimensional model can be viewed as special cases of the multi-unidimensional model. Moreover, introduce an unconstrained covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = [\sigma_{vv'}]_{m\times m}$, so that the constrained covariance matrix $\mathbf{P}$ can be readily transformed from $\boldsymbol{\Sigma}$ using

$$\rho_{st} = \frac{\sigma_{st}}{\sqrt{\sigma_{ss}\sigma_{tt}}}, \qquad s \neq t. \tag{3}$$

A noninformative prior can be assumed for $\boldsymbol{\Sigma}$ so that $p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{m+1}{2}}$ (Lee 1995).

Hence, with prior distributions assumed for $\boldsymbol{\xi}_{vj}$, where $\boldsymbol{\xi}_{vj} = (\alpha_{vj}, \beta_{vj})'$, the joint posterior distribution of ($\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\mathbf{Z}$, $\boldsymbol{\Sigma}$) is

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Z}, \boldsymbol{\Sigma}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{Z})p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi})p(\boldsymbol{\xi})p(\boldsymbol{\theta}|\mathbf{P})p(\boldsymbol{\Sigma}), \tag{4}$$

where

$$f(\mathbf{y}|\mathbf{Z}) = \prod_{v=1}^{m}\prod_{i=1}^{n}\prod_{j=1}^{k_v} p_{vij}^{y_{vij}}(1 - p_{vij})^{1-y_{vij}} \tag{5}$$

is the likelihood function, with p_{vij} being the probability function for the multi-unidimensional model as defined in (1).

The implementation of the Gibbs sampling procedure thus involves four of the sampling processes, namely, a sampling of the augmented Z parameters from

$$Z_{vij}|\bullet \sim \begin{cases} N_{(0,\infty)}(\eta_{vij}, 1), & \text{if} \quad y_{vij} = 1 \\ N_{(-\infty,0)}(\eta_{vij}, 1), & \text{if} \quad y_{vij} = 0 \end{cases}; \tag{6}$$

a sampling of person traits $\boldsymbol{\theta}$ from

$$\boldsymbol{\theta}_i|\bullet \sim N_m((\mathbf{A}'\mathbf{A} + \mathbf{P})^{-1}\mathbf{A}'\mathbf{B}, (\mathbf{A}'\mathbf{A} + \mathbf{P})^{-1}), \tag{7}$$

where

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{\alpha}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\alpha}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\alpha}_m \end{bmatrix}_{K\times m} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{Z}_{1i} + \boldsymbol{\beta}_1 \\ \mathbf{Z}_{2i} + \boldsymbol{\beta}_2 \\ \vdots \\ \mathbf{Z}_{mi} + \boldsymbol{\beta}_m \end{bmatrix}_{K\times 1},$$

in which $\boldsymbol{\alpha}_v = (\alpha_{v1}, ..., \alpha_{vk_v})'$, $\mathbf{Z}_{vi} = (Z_{vi1}, ..., Z_{vk_v})'$, $\boldsymbol{\beta}_v = (\beta_{v1}, ..., \beta_{vk_v})'$; a sampling of the item parameters $\boldsymbol{\xi}$ from

$$\boldsymbol{\xi}_{vj}|\bullet \sim N_2((\mathbf{x}_v'\mathbf{x}_v)^{-1}\mathbf{x}_v'\mathbf{Z}_{vj}, (\mathbf{x}_v'\mathbf{x}_v)^{-1})I(\alpha_{vj} > 0), \tag{8}$$

where $\mathbf{x}_v = [\boldsymbol{\theta}_v, -1]$, assuming noninformative uniform priors $\alpha_{vj} > 0$ and $\beta_{vj} \propto 1$, or from

$$\boldsymbol{\xi}_{vj}|\bullet \sim N_2((\mathbf{x}_v'\mathbf{x}_v + \boldsymbol{\Sigma}_{\boldsymbol{\xi}_v}^{-1})^{-1}(\mathbf{x}_v'\mathbf{Z}_{vj} + \boldsymbol{\Sigma}_{\boldsymbol{\xi}_v}^{-1}\mu_{\xi_v}), (\mathbf{x}_v'\mathbf{x}_v + \boldsymbol{\Sigma}_{\boldsymbol{\xi}_v}^{-1})^{-1})I(\alpha_{vj} > 0), \tag{9}$$

where $\boldsymbol{\mu}_{\boldsymbol{\xi}_v} = (\mu_{\alpha_v}, \mu_{\beta_v})'$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_v} = \begin{pmatrix} \sigma_{\alpha_v}^2 & 0 \\ 0 & \sigma_{\beta_v}^2 \end{pmatrix}$, assuming conjugate normal priors $\alpha_{vj} \sim N_{(0,\infty)}(\mu_{\alpha_v}, \sigma_{\alpha_v}^2)$, $\beta_{vj} \sim N(\mu_{\beta_v}, \sigma_{\beta_v}^2)$; and a sampling of the unconstrained covariance matrix $\boldsymbol{\Sigma}$ from

$$\boldsymbol{\Sigma}|\bullet \sim W^{-1}(\mathbf{S}^{-1}, n) \tag{10}$$

(an inverse Wishart distribution), where $\mathbf{S} = \sum_{i=1}^{n} (\mathbf{C}\boldsymbol{\theta}_i)(\mathbf{C}\boldsymbol{\theta}_i)'$, in which

$$\mathbf{C} = \begin{bmatrix} (\prod_{j=1}^{k_1} \alpha_{1j})^{1/k_1} & 0 & \cdots & 0 \\ 0 & (\prod_{j=1}^{k_2} \alpha_{2j})^{1/k_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & (\prod_{j=1}^{k_m} \alpha_{mj})^{1/k_m} \end{bmatrix}_{m \times m}$$

(see Lee 1995, for a detailed derivation of the full conditional distributions). From each sampled $\boldsymbol{\Sigma}$, the constrained covaraince matrix $\mathbf{P}$ can be obtained using (3). Hence, with starting values $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\xi}^{(0)}$, and $\mathbf{P}^{(0)}$, observations ($\mathbf{Z}^{(\ell)}$, $\boldsymbol{\theta}^{(\ell)}$, $\boldsymbol{\xi}^{(\ell)}$, $\boldsymbol{\Sigma}^{(\ell)}$, $\mathbf{P}^{(\ell)}$) can be drawn or tranformed iteratively from (6), (7), (8), (10) and (3) (or from (6), (7), (9), (10) and (3)), respectively.

This iterative process continues for a sufficient number of samples after the posterior distributions reach stationarity (a phase known as burn-in). The posterior means of all the samples collected after the burn-in stage are considered as estimates of the true model parameters ($\boldsymbol{\xi}$, $\boldsymbol{\theta}$) and the hyperparameter ($\mathbf{P}$). Similarly, their posterior standard deviations are used to describe the statistical uncertainty. However, Monte Carlo standard errors cannot be calculated using the sample standard deviations because subsequent samples in each Markov chain are autocorrelated (e.g., Patz and Junker 1999b). Among the standard methods for estimating them (Ripley 1987), batching is said to be a crude but effective method (Verdinelli and Wasserman 1995) and hence is considered in this paper. Here, with a long chain of samples being separated into contiguous batches of equal length, the Monte Carlo standard error for each parameter is then estimated to be the standard deviation of these batch means. The Monte Carlo standard error of the estimate is hence a ratio of the Monte Carlo standard error and the square root of the number of batches. More sophisticated methods for estimating standard errors can be found in Gelman and Rubin (1992).

3. Bayesian model choice or checking

In Bayesian statistics, the adequacy of the fit of a given model is evaluated using several model choice or checking techniques, among which, Bayesian deviance and posterior predictive model checks are considered and briefly illustrated.

3.1. Bayesian deviance

The Bayesian deviance information criterion (DIC; Spiegelhalter *et al.* 1998) is based on the posterior distribution of the deviance. This criterion is defined as DIC $= \overline{D} + p_D$, where $\overline{D} = E(-2\log L(\mathbf{y}|\boldsymbol{\vartheta}))$ is the posterior expectation of the deviance (with $L(\mathbf{y}|\boldsymbol{\vartheta})$ being the model likelihood function, where $\boldsymbol{\vartheta}$ denotes all model parameters) and $p_D = \overline{D} - D(\bar{\boldsymbol{\vartheta}})$ is the effective number of parameters (Carlin and Louis 2000). Further, $D(\bar{\boldsymbol{\vartheta}}) = -2\log(L(\mathbf{y}|\bar{\boldsymbol{\vartheta}}))$, where $\bar{\boldsymbol{\vartheta}}$ is the posterior mean. To compute Bayesian DIC, MCMC samples of the parameters, $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(G)}$, can be drawn using the Gibbs sampler, then $\overline{D} = 1/G(-2\log\prod_{g=1}^{G} L(\mathbf{y}|\boldsymbol{\vartheta}^{(g)}))$. Small values of the deviance suggest a better-fitting model. Generally, more complicated models tend to provide better fit. Hence, penalizing for the number of parameters (p_D) makes DIC a more reasonable measure to use.

3.2. Posterior predictive model checks

The posterior predictive model checking (PPMC; Rubin 1984) method provides a popular Bayesian model checking technique that is intuitively appealing, simple to implement, and easy to interpret (Sinharay and Stern 2003). The basic idea is to draw replicated data $\mathbf{y}^{\text{rep}}$ from its posterior predictive distribution $p(\mathbf{y}^{\text{rep}}|\mathbf{y}) = \int p(\mathbf{y}^{\text{rep}}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\mathbf{y})d\boldsymbol{\vartheta}$, and compare them to the observed data $\mathbf{y}$. If the model fits, then replicated data generated under the model should look similar to the observed data. A test statistic known as the discrepancy measure $T(\mathbf{y}, \boldsymbol{\vartheta})$ has to be chosen to define the discrepancy between the model and the data. For each $\mathbf{y}^{\text{rep}}$ drawn from the predictive distribution, the realized discrepancy $T(\mathbf{y}) = T(\mathbf{y}, \boldsymbol{\vartheta})$ can be compared with the predictive discrepancy $T(\mathbf{y}^{\text{rep}}) = T(\mathbf{y}^{\text{rep}}, \boldsymbol{\vartheta})$ by plotting the pairs on a scatter plot. Alternatively, one can obtain a quantitative measure of lack of fit by calculating the tail-area probability or the PPP value (Sinharay *et al.* 2006), $P(T(\mathbf{y}^{\text{rep}}) \geq T(\mathbf{y})|\mathbf{y}) = \int_{T(\mathbf{y}^{\text{rep}})\geq T(\mathbf{y})} p(\mathbf{y}^{\text{rep}}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\mathbf{y})d\mathbf{y}^{\text{rep}}d\boldsymbol{\vartheta}$.

To implement the method, one draws G samples from the posterior distribution of $\boldsymbol{\vartheta}$ using Gibbs sampling. Then for each simulated $\boldsymbol{\vartheta}^{(g)}$, a $\mathbf{y}^{\text{rep}(g)}$ can be drawn from the predictive distribution so there are G draws from the joint posterior distribution $p(\mathbf{y}^{\text{rep}}, \boldsymbol{\vartheta}|\mathbf{y})$. The predictive test statistic $T(\mathbf{y}^{\text{rep}(g)})$ and the realized test statistic $T(\mathbf{y})$ are computed and subsequently compared to provide graphical and numerical evidence about model inadequacy. Specifically, the proportion of the G simulated samples for which the replicated data could be more extreme than the observed data, i.e., $1/G\sum_{g=1}^{G} I(T(\mathbf{y}^{\text{rep}(g)}) \geq T(\mathbf{y}))$, provides an estimate of the PPP value. Extreme PPP values (close to 0 or 1) indicate model misfit. It has to be noted that although both are defined as tail-area probabilities, the PPP value has to be differentiated from the traditional hypothesis-testing p value in that the posterior predictive checking approach does not perform a hypothesis test (Gelman *et al.* 1996). The choice of the discrepancy measure is critical in implementing this method. Sinharay *et al.* (2006), after evaluating a number of discrepancy measures for assessing IRT models, concluded that the odds ratio for measuring associations among item pairs, $T(\mathbf{y}) = OR_{ij} = \frac{n_{11}n_{00}}{n_{01}n_{10}}$, is powerful for

detecting lack of fit when the model assumption is violated. Hence, this measure is adopted as the test statistic for the PPMC method in this paper.

4. Package IRTmu2no

The package **IRTmu2no** contains two major user-callable routines. A function for generating binary response data using the 2PNO multi-unidimensional IRT model titled, `simmu2no`, and a function that implements MCMC to obtain posterior samples, estimates, convergence statistics, or model choice/checking statistics, `gsmu2no`.

The function `simmu2no` has input arguments `n`, `kv`, `r`, and `iparm` for the number of respondents, the number of items in each subtest, the actual population correlation matrix for person latent traits, and user-specified item parameters, respectively. The optional `iparm` argument allows the user the choice to input item parameters for the model, or randomly generate them from uniform distributions so that $\alpha_{vj} \sim U(0, 2)$ and $\beta_{vj} \sim U(-2, 2)$. The user can further choose to store the simulated person (`theta`) and item (`item`) parameters.

With the required user-input binary response data (`y`) and the optional number of items in each subtest (`kv`), the function `gsmu2no` initially reads in starting values for person and item parameters (`th0`, `item0`) and the person hyperparameter (`sigma0`), or sets them to be $\theta_{vi}^{(0)} = 0$, $\alpha_{vj}^{(0)} = 2$, $\beta_{vj}^{(0)} = -\Phi^{-1}(\sum_i y_{vij}/n)\sqrt{5}$ (following Albert 1992) and $\mathbf{P} = \mathbf{I}$. It then implements the Gibbs sampler for the 2PNO multi-unidimensional IRT model by iteratively drawing random samples for the parameters from their respective full conditional distributions. The prior distributions for the item parameters can be noninformative (`flat = 1`, default) or informative (`flat = 0`). In the latter case, the user can specify any values of interest or use the default values, namely, $\mu_{\alpha_v} = 0$ and $\sigma^2_{\alpha_v} = 1$ for α_{vj} (`aprior`), and $\mu_{\beta_v} = 0$ and $\sigma^2_{\beta_v} = 1$ for β_{vj} (`gprior`). It is noted that the prior location and scale parameters for α_{vj} or β_{vj} can be set to be different across the m subtests. The algorithm continues until all the (`kk`) samples are simulated, with the early burn-in samples (`burnin`) being discarded, where `kk` and `burnin` can be $10,000$ and `kk`$/2$ (default) or any values of interest. It then computes the posterior estimates, posterior standard deviations, and Monte Carlo standard errors of the person (`pparm`), item (`iparm`) or intertrait correlation (`rho`) estimates. Posterior samples of these parameters can also be stored (`samples`) for further analysis.

In addition to Monte Carlo standard errors, convergence can be evaluated using the Gelman-Rubin R statistic (Gelman *et al.* 2004) for each model parameter. The usual practice is using multiple Markov chains from different starting points. Alternatively, a single chain can be divided into sub-chains so that convergence is assessed by comparing the between and within sub-chain variance. Since a single chain is less wasteful in the number of iterations needed, the latter approach is adopted to compute the R statistic (`gr`) with `gsmu2no`. The Bayesian deviance estimates, including $\overline{D}$, $D(\bar{\vartheta})$, p_D and DIC, can be obtained (`deviance`) to measure the relative fit of a model. Moreover, the PPMC method can be adopted using the odds ratio as the discrepancy measure so that PPP values (`ppmc`) are obtained to evaluate model misfit. Extreme PPP values can be further plotted using the function `ppmcplt`, in which the threshold (`crit`) can be 0.01 (default) or any level of interest so that PPP values larger than $1 -$ `crit`$/2$ are denoted using the right triangle sign and those smaller than `crit`$/2$ are denoted using the left triangle sign. The functions' input and output arguments are completely specified in the m-files.

5. Illustrative examples

To demonstrate the use of the **IRTmu2no** package, simulated and real data examples are provided in this section to illustrate item parameter recovery as well as model comparisons. The code to reproduce the results of each example is provided in the m-file `v28i10.m`, which may also serve as a guide for the user of the Gibbs sampler with the 2PNO multi-unidimensional IRT model.

5.1. Parameter recovery

For parameter recovery, tests with two subtests were considered so that the first half measured one latent trait and the second half measured another. Three 1000-by-18 (i.e., $n = 1000$, $m = 2$, $k_1 = 9$, $k_2 = 9$, and $K = 18$) dichotomous data matrices were simulated from the 2PNO multi-unidimensional model where the actual correlation (ρ_{12}) between the two distinct traits (θ_{1i}, θ_{2i}) was set to be 0.2, 0.5, and 0.8, respectively. The item parameters were taken from Li and Schafer (2005, p.11), which are shown in the first column of Tables 1, 2 and 3. Gibbs sampling was subsequently implemented to recover model parameters assuming the noninformative or informative prior distributions described previously. The posterior means and standard deviations of item parameters (α_v, β_v) as well as the intertrait correlation hyperparameter (ρ_{12}), together with their Monte Carlo standard errors of estimates and Gelman-Rubin R statistics were obtained and are displayed in the rest of the tables. The overall estimation accuracy was also evaluated using the average square error between the actual and estimated parameters, which is shown at the bottom of each table.

The Gelman-Rubin R statistic provides a numerical measure for assessing convergence for each model parameter. With values close to 1, it is determined that in the implementations of the Gibbs sampler, Markov chains reached stationarity with a run length of $10,000$ iterations and a burn-in period of $5,000$ iterations. Convergence can also be monitored visually using time series graphs of the simulated sequence, such as the trace plot, the running mean plot, and the autocorrelation plot shown in Figure 1 for one item. The autocorrelations between successive parameter draws became negligible at lags greater than 70. According to Geyer (1992), burn-in for a single chain should not take longer than the number of iterations required to achieve negligible autocorrelations. Indeed, the trace plot and the running mean plot both suggest that $5,000$ iterations were long enough for the chains to converge.

The results summarized in the three tables indicate that the item parameters as well as the intertrait correlation hyperparameter were estimated with enough accuracy, suggesting that the multi-unidimensional IRT model performs well in various test situations where the distinct latent dimensions have a low, medium or high correlation. In addition, the two sets of posterior estimates, resulted from different prior distributions, differ only slightly from each other, signifying that the posterior estimates are not sensitive to the choice of noninformative or informative priors for the slope and intercept parameters. This point is further supported by the small difference between the average square errors.

5.2. Model comparison

To illustrate the use of the Bayesian model choice or checking techniques for evaluating the relative fit of a model, three IRT models were considered in model comparisons, namely, the multi-unidimensional model, the unidimensional model, and a constrained multi-unidimensional

	Noninformative priors				Informative priors			
True	Estimate	SD	MCSE	R	Estimate	SD	MCSE	R
α_1								
0.621	0.6991	0.0653	0.0028	1.011	0.6925	0.0667	0.0029	1.007
1.190	1.5481	0.1502	0.0178	1.001	1.4759	0.1447	0.0151	1.097
0.778	0.8598	0.0715	0.0036	1.009	0.8579	0.0725	0.0030	1.001
1.627	1.6178	0.1725	0.0176	1.035	1.5802	0.1696	0.0150	1.023
1.056	1.0243	0.1181	0.0065	1.017	0.9909	0.1104	0.0128	1.055
1.411	1.2984	0.1239	0.0139	1.056	1.2588	0.1080	0.0077	1.011
0.482	0.4112	0.0561	0.0021	1.007	0.4035	0.0588	0.0026	1.011
0.963	1.0544	0.0993	0.0075	1.038	1.0494	0.1053	0.0073	1.040
0.700	0.7013	0.0666	0.0016	1.003	0.6848	0.0708	0.0034	1.012
α_2								
0.361	0.4468	0.0813	0.0037	1.001	0.4469	0.0844	0.0032	1.008
0.515	0.4779	0.0705	0.0027	1.008	0.4741	0.0682	0.0020	1.004
1.078	1.0106	0.1097	0.0074	1.012	0.9883	0.1123	0.0103	1.056
0.809	0.6866	0.0714	0.0030	1.017	0.6803	0.0734	0.0017	1.002
0.433	0.2601	0.0740	0.0041	1.022	0.2495	0.0720	0.0032	1.019
1.069	0.9326	0.0993	0.0062	1.032	0.9193	0.0890	0.0041	1.011
0.818	0.8617	0.0971	0.0049	1.018	0.8551	0.0963	0.0063	1.010
0.811	0.7444	0.0784	0.0031	1.008	0.7491	0.0730	0.0033	1.005
0.786	0.7843	0.0805	0.0033	1.005	0.7892	0.0836	0.0043	1.017
β_1								
0.390	0.3210	0.0494	0.0023	1.007	0.3190	0.0485	0.0020	1.011
−1.061	−1.1342	0.1053	0.0120	1.006	−1.1039	0.0979	0.0079	1.047
0.294	0.3415	0.0545	0.0021	1.002	0.3399	0.0530	0.0023	1.009
−0.760	−0.7564	0.0913	0.0050	1.011	−0.7543	0.0898	0.0070	1.027
1.533	1.5679	0.1059	0.0055	1.017	1.5397	0.1046	0.0102	1.037
0.873	0.7412	0.0756	0.0046	1.035	0.7201	0.0707	0.0040	1.015
0.878	0.8669	0.0497	0.0012	1.000	0.8624	0.0507	0.0020	1.014
1.174	1.1830	0.0867	0.0065	1.015	1.1815	0.0870	0.0053	1.029
0.912	0.9024	0.0589	0.0024	1.003	0.8930	0.0577	0.0021	1.007
β_2								
1.475	1.4746	0.0690	0.0030	1.001	1.4743	0.0775	0.0022	1.004
0.851	0.9314	0.0561	0.0023	1.008	0.9249	0.0546	0.0014	1.002
−0.678	−0.6284	0.0662	0.0043	1.015	−0.6249	0.0671	0.0052	1.019
0.396	0.4433	0.0498	0.0014	1.002	0.4415	0.0501	0.0018	1.000
1.545	1.4600	0.0625	0.0032	1.028	1.4503	0.0631	0.0028	1.014
0.381	0.3711	0.0574	0.0028	1.014	0.3644	0.0564	0.0029	1.001
0.845	0.9050	0.0684	0.0033	1.011	0.8962	0.0683	0.0049	1.006
−0.332	−0.2846	0.0501	0.0020	1.010	−0.2871	0.0497	0.0017	1.003
−0.293	−0.2158	0.0517	0.0017	1.002	−0.2194	0.0517	0.0024	1.002
ρ_{12}								
0.200	0.1601	0.0428	0.0009	1.001	0.1634	0.0427	0.0013	1.005
	Average square error = 0.0085				Average square error = 0.0080			

Table 1: Posterior estimate, standard deviation (SD), Monte Carlo standard error of the estimate (MCSE) and Gelman-Rubin R statistic for α_v, β_v, and ρ_{12} when the true intertrait correlation is 0.2 (chain length = 10,000, burn-in = 5,000).

	Noninformative priors				Informative priors			
True	Estimate	SD	MCSE	R	Estimate	SD	MCSE	R
α_1								
0.621	0.5338	0.0560	0.0012	1.002	0.5306	0.0558	0.0016	1.005
1.190	1.5487	0.1640	0.0157	1.015	1.5150	0.1620	0.0185	1.038
0.778	0.8885	0.0743	0.0036	1.003	0.8801	0.0757	0.0041	1.007
1.627	1.6462	0.1580	0.0171	1.093	1.5997	0.1566	0.0197	1.123
1.056	1.3140	0.1705	0.0290	1.234	1.2438	0.1364	0.0150	1.059
1.411	1.4317	0.1231	0.0076	1.011	1.4305	0.1342	0.0102	1.025
0.482	0.5202	0.0622	0.0027	1.012	0.5206	0.0615	0.0022	1.004
0.963	0.9577	0.0940	0.0080	1.029	0.9347	0.0898	0.0062	1.038
0.700	0.6648	0.0680	0.0036	1.015	0.6687	0.0667	0.0040	1.016
α_2								
0.361	0.3992	0.0710	0.0037	1.016	0.3912	0.0726	0.0030	0.999
0.515	0.5049	0.0629	0.0025	1.009	0.4976	0.0611	0.0023	1.004
1.078	1.0407	0.1018	0.0082	1.027	1.0429	0.0968	0.0041	1.011
0.809	0.8830	0.0821	0.0029	1.008	0.8704	0.0821	0.0041	1.016
0.433	0.4893	0.0883	0.0050	1.022	0.4758	0.0829	0.0045	1.001
1.069	0.9864	0.0935	0.0068	1.008	0.9768	0.0889	0.0049	1.008
0.818	0.7804	0.0799	0.0039	1.015	0.7751	0.0799	0.0042	1.028
0.811	0.8136	0.0772	0.0042	1.020	0.8197	0.0784	0.0037	1.022
0.786	0.8622	0.0815	0.0025	1.001	0.8477	0.0825	0.0047	1.017
β_1								
0.390	0.3116	0.0472	0.0017	1.002	0.3033	0.0452	0.0012	1.005
−1.061	−1.3245	0.1222	0.0107	1.021	−1.3217	0.1255	0.0134	1.032
0.294	0.3163	0.0546	0.0027	1.012	0.3056	0.0541	0.0027	1.019
−0.760	−0.8266	0.0942	0.0074	1.052	−0.8345	0.0941	0.0097	1.061
1.533	1.7951	0.1658	0.0275	1.240	1.7188	0.1331	0.0141	1.050
0.873	0.7578	0.0808	0.0032	1.006	0.7439	0.0825	0.0054	1.024
0.878	0.9159	0.0549	0.0019	1.010	0.9080	0.0542	0.0021	1.006
1.174	1.2053	0.0790	0.0063	1.036	1.1772	0.0751	0.0046	1.033
0.912	0.9378	0.0574	0.0025	1.018	0.9322	0.0595	0.0035	1.020
β_2								
1.475	1.4202	0.0633	0.0033	1.013	1.4072	0.0635	0.0030	1.002
0.851	0.7933	0.0514	0.0016	1.008	0.7867	0.0502	0.0026	1.010
−0.678	−0.6516	0.0667	0.0040	1.009	−0.6590	0.0658	0.0026	1.004
0.396	0.4066	0.0567	0.0024	1.007	0.3957	0.0551	0.0025	1.009
1.545	1.6080	0.0826	0.0034	1.009	1.5896	0.0789	0.0043	1.017
0.381	0.3709	0.0588	0.0027	1.004	0.3631	0.0569	0.0025	1.010
0.845	0.8122	0.0614	0.0032	1.014	0.8031	0.0596	0.0030	1.017
−0.332	−0.3540	0.0534	0.0016	1.000	−0.3593	0.0529	0.0023	1.009
−0.293	−0.3579	0.0533	0.0015	1.002	−0.3635	0.0529	0.0026	1.005
ρ_{12}								
0.500	0.5071	0.0363	0.0014	1.011	0.5096	0.0360	0.0019	1.006
	Average square error = 0.0115				Average square error = 0.0092			

Table 2: Posterior estimate, standard deviation (SD), Monte Carlo standard error of the estimate (MCSE) and Gelman-Rubin R statistic for α_v, β_v, and ρ_{12} when the true intertrait correlation is 0.5 (chain length = 10,000, burn-in = 5,000).

	Noninformative priors				Informative priors			
True	Estimate	SD	MCSE	R	Estimate	SD	MCSE	R
α_1								
0.621	0.6709	0.0630	0.0025	1.002	0.6677	0.0611	0.0033	1.013
1.190	1.1436	0.1045	0.0078	1.031	1.1193	0.1013	0.0056	1.003
0.778	0.7918	0.0658	0.0022	1.004	0.7880	0.0663	0.0018	1.006
1.627	2.0085	0.2334	0.0297	1.031	1.8873	0.1797	0.0271	1.219
1.056	1.3649	0.1446	0.0165	1.073	1.3409	0.1410	0.0153	1.067
1.411	1.3799	0.1233	0.0102	1.028	1.3695	0.1178	0.0108	1.079
0.482	0.5271	0.0613	0.0032	1.016	0.5249	0.0606	0.0028	1.014
0.963	1.0194	0.0876	0.0046	1.003	1.0187	0.0941	0.0056	1.003
0.700	0.6432	0.0648	0.0013	1.000	0.6362	0.0650	0.0033	1.000
α_2								
0.361	0.3180	0.0674	0.0036	1.007	0.3125	0.0693	0.0037	1.017
0.515	0.4071	0.0570	0.0021	1.001	0.4066	0.0548	0.0020	1.005
1.078	1.2643	0.1252	0.0098	1.053	1.2456	0.1248	0.0141	1.083
0.809	0.7838	0.0672	0.0020	1.006	0.7770	0.0699	0.0030	1.003
0.433	0.4327	0.0772	0.0043	1.014	0.4229	0.0770	0.0038	1.006
1.069	1.1157	0.0980	0.0062	1.013	1.1131	0.0939	0.0073	1.018
0.818	0.9498	0.0880	0.0044	1.010	0.9299	0.0816	0.0059	1.005
0.811	0.8708	0.0778	0.0035	1.005	0.8558	0.0777	0.0032	1.008
0.786	0.8810	0.0784	0.0036	1.008	0.8651	0.0732	0.0049	1.022
β_1								
0.390	0.3968	0.0479	0.0011	1.001	0.3926	0.0496	0.0021	1.012
−1.061	−1.0121	0.0805	0.0044	1.007	−1.0059	0.0781	0.0052	1.014
0.294	0.2945	0.0505	0.0024	1.008	0.2906	0.0508	0.0024	1.014
−0.760	−0.8590	0.1147	0.0106	1.009	−0.8310	0.1021	0.0075	1.039
1.533	1.7961	0.1346	0.0148	1.078	1.7673	0.1338	0.0155	1.094
0.873	0.9102	0.0825	0.0056	1.007	0.9018	0.0845	0.0079	1.074
0.878	0.8460	0.0535	0.0014	1.002	0.8388	0.0527	0.0030	1.020
1.174	1.1675	0.0737	0.0037	1.003	1.1644	0.0811	0.0043	1.007
0.912	0.8136	0.0545	0.0016	1.000	0.8076	0.0552	0.0031	1.014
β_2								
1.475	1.4382	0.0628	0.0031	1.010	1.4285	0.0631	0.0025	1.007
0.851	0.7391	0.0491	0.0016	1.003	0.7369	0.0488	0.0014	1.003
−0.678	−0.7489	0.0772	0.0047	1.023	−0.7461	0.0770	0.0065	1.053
0.396	0.3360	0.0507	0.0015	1.003	0.3291	0.0515	0.0015	1.005
1.545	1.5866	0.0737	0.0044	1.022	1.5721	0.0709	0.0035	1.007
0.381	0.3161	0.0591	0.0030	1.011	0.3127	0.0580	0.0026	1.014
0.845	0.8988	0.0678	0.0034	1.010	0.8851	0.0648	0.0031	1.008
−0.332	−0.4663	0.0553	0.0018	1.003	−0.4712	0.0561	0.0022	1.011
−0.293	−0.2833	0.0531	0.0025	1.004	−0.2850	0.0513	0.0018	1.003
ρ_{12}								
0.800	0.7804	0.0286	0.0016	1.026	0.7885	0.0297	0.0015	1.005
	Average square error = 0.0130				Average square error = 0.0097			

Table 3: Posterior estimate, standard deviation (SD), Monte Carlo standard error of the estimate (MCSE) and Gelman-Rubin R statistic for α_v, β_v, and ρ_{12} when the true intertrait correlation is 0.8 (chain length = 10,000, burn-in = 5,000).

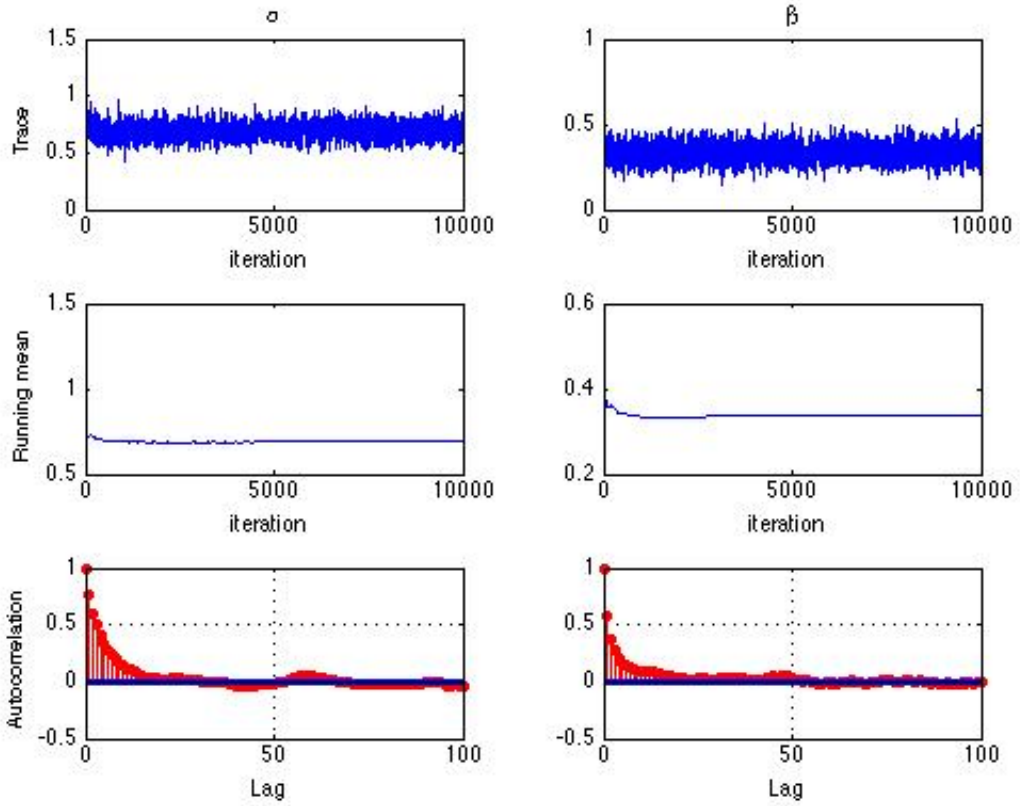


Figure 1: Trace plots (upper), running mean plots (middle), and autocorrelation plots (lower) of 10,000 draws of α_v and β_v for an item in one implementation.

model where $\boldsymbol{\theta}_i \sim N_m(\mathbf{0}, \mathbf{I})$ (note that this model is equivalent to fitting the unidimensional model separately for each subtest because the latent traits are assumed to be uncorrelated, and hence it is denoted using "unidimensional model (2)" in the following discussion). As illustrated in Section 2, the latter two models can be viewed as special cases of the former where the latent traits have a perfect or a zero correlation. Three 1000-by-30 dichotomous data matrices were simulated from the 2PNO multi-unidimensional IRT model so that 15 items measured θ_1 and another 15 items measured θ_2 with their intertrait correlation ρ_{12} being 0, 0.5, and 1, respectively. It is noted that although a 0 or 1 correlation is rarely observed in practice, it was considered here to illustrate extreme situations where the unidimensional model for the overall test or that for individual subtests would not be adequate. To generate the data matrices, item parameters were randomly drawn from the uniform distributions described in Section 4. The Gibbs sampler assuming noninformative priors for the item parameters was subsequently implemented so that 10,000 samples were simulated with the first 5,000 set to be burn-in. The Gelman-Rubin R statistics suggest that the chains converged to their stationary distributions within 10,000 iterations. Hence, the Bayesian deviance estimates, including $\overline{D}$, $D(\bar{\vartheta})$, p_D and DIC, were obtained from each implementation and are displayed in Table 4. It is clear from the table that after accounting for model complexity, the Bayesian DIC (column 5) estimates pointed to the more general multi-unidimensional model, even in

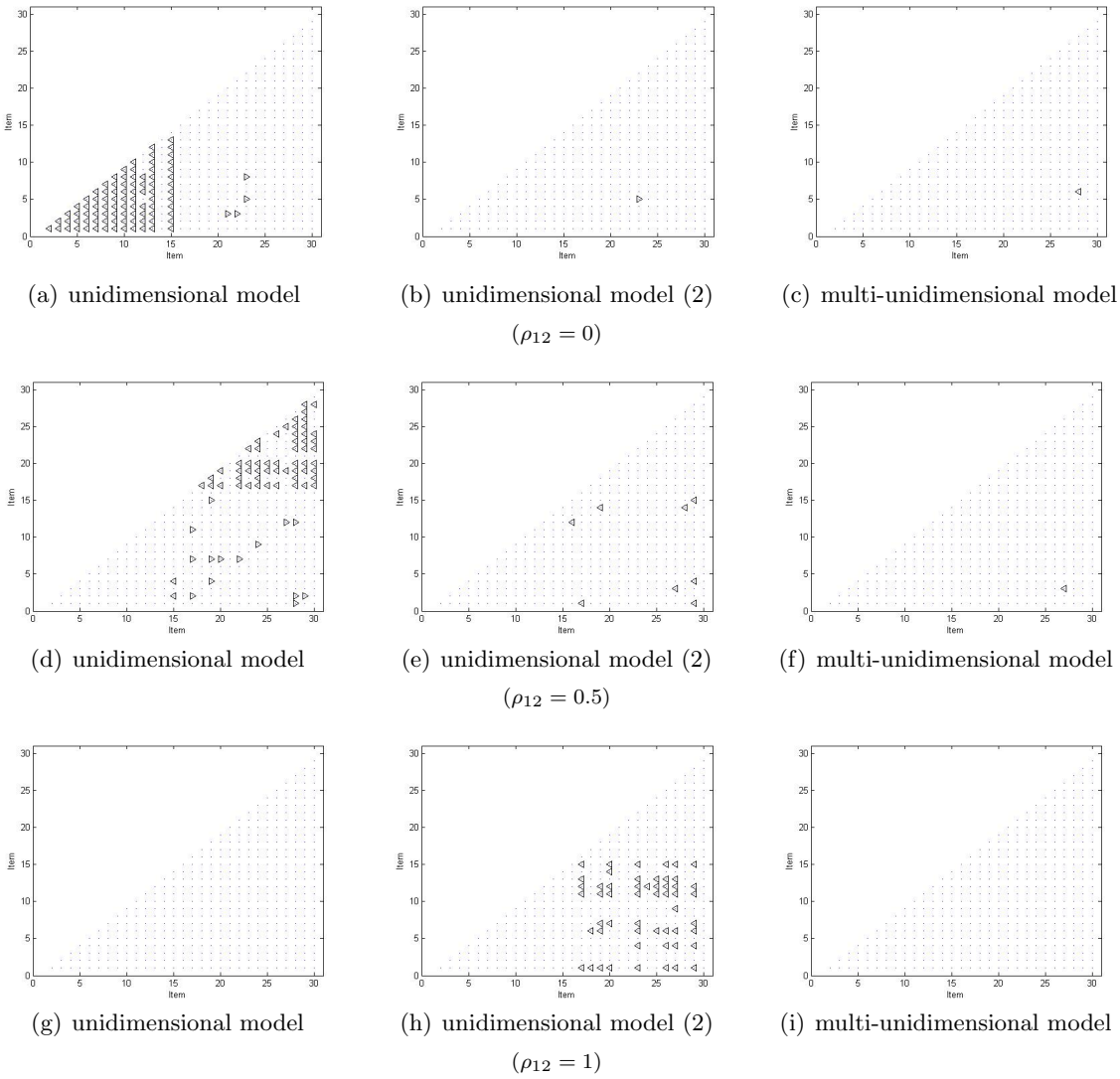
	$\overline{D}$	$D(\bar{\vartheta})$	p_D	DIC
ρ_{12}=0				
unidimensional model	27794.3066	26873.0312	921.2753	28715.5819
unidimensional model (2)	22706.8202	20924.6603	1782.1598	24488.9800
multi-unidimensional model	22695.5202	20924.2061	1771.3141	24466.8343
ρ_{12}=.5				
unidimensional model	24472.8049	23541.7085	931.0964	25403.9013
unidimensional model (2)	21772.1656	20048.5712	1723.5945	23495.7601
multi-unidimensional model	21784.4596	20120.6616	1663.7980	23448.2576
ρ_{12}=1				
unidimensional model	24122.5643	23161.4451	961.1192	25083.6835
unidimensional model (2)	24089.3970	22389.2960	1700.1010	25789.4979
multi-unidimensional model	24004.6142	22927.7444	1076.8698	25081.4839

Table 4: Bayesian deviance estimates for multi-unidimensional ($\rho_{12} = 0$, $\rho_{12} = 0.5$) or unidimensional ($\rho_{12} = 1$) data fitted with the three IRT models.

situations where the test measured one common latent trait ($\rho_{12} = 1$) or two completely different traits ($\rho_{12} = 0$). It has to be noted that when data were clearly unidimensional, the difference between the DIC estimates for the unidimensional and the multi-unidimensional models was rather trivial. Thus, one may argue that these two models were essentially similar given that the small difference might not make a practical significance.

PPMC was also implemented to obtain PPP values for the three IRT models where odds ratios were used as the discrepancy measure. Graphical representations of extreme PPP values are shown in Figure 2, where the upper diagonal is left blank for each plot due to symmetry. Here, with a threshold of 0.01, the plots indicate that the multi-unidimensional model had consistently fewer number of extreme predicted odds ratios and hence provided a description no worse, if not better than the other two models in all the three situations considered. On the other hand, the larger number of extreme PPP values clearly suggest the lack of fit of the unidimensinal model when $\rho_{12} = 0$ (see Figure 2(a)), and that of the unidimensional model (2) when $\rho_{12} = 1$ (see Figure 2(h)), which is consistent with the previous result using DIC. Interestingly, when $\rho_{12} = 0.5$, the unidimensional model (2) outperformed the unidimensional model (see Table 4 and Figures 2(d), 2(e)), suggesting that although fitting the unidimensional model to individual subtests is not the best approach, it tends to be more appropriate than fitting the unidimensional model to the overall test when the latent traits are believed to be moderately correlated.

Consistent with findings from Sheng and Wikle (2007), the model comparison results based on the Bayesian deviance and PPMC criteria suggest that the multi-unidimensional IRT model is applicable in a wider range of test situations where the unidimensional model for the overall test or for individual subtests is not appropriate, and is even robust to the true process being unidimensional. Hence, allowing separate inferences made about a person's multiple traits while modeling their underlying structure, the multi-unidimensional model provides a better and more flexible way to represent test data not realized in the conventional IRT model.



(a) unidimensional model (b) unidimensional model (2) ($\rho_{12} = 0$) (c) multi-unidimensional model

(d) unidimensional model (e) unidimensional model (2) ($\rho_{12} = 0.5$) (f) multi-unidimensional model

(g) unidimensional model (h) unidimensional model (2) ($\rho_{12} = 1$) (i) multi-unidimensional model

Figure 2: Plots of PPP values for odds ratios for multi-unidimensional ($\rho_{12} = 0$, $\rho_{12} = 0.5$) or unidimensional ($\rho_{12} = 1$) data fitted with the three IRT models. (Triangles represent extreme PPP values, where values smaller than 0.005 are denoted using left triangle signs and values larger than 0.995 are denoted using right triangle signs.)

5.3. Empirical example

A subset of the College Basic Academic Subjects Examination (CBASE; Osterlind 1997) English data was further used to illustrate the Bayesian approach for model choice or model checking. The data contains binary responses of $1,200$ independent college students to a total of 41 English multiple-choice items. The English test is further organized into levels of increasing specificity by two subtests, namely, writing and reading, so that 16 items are in one subtest and 25 are in the other. One may note that the nature of the test limits the candidate models to be either unidimensional or multi-unidimensional. Model comparison is consequently necessary for establishing the model that provides a relatively better represen-

	$\overline{D}$	$D(\bar{\vartheta})$	p_D	DIC
unidimensional model	53973.9122	52915.7493	1058.1630	55032.0752
unidimensional model (2)	53500.2910	51748.5462	1751.7448	55252.0358
multi-unidimensional model	53486.5558	52068.6752	1417.8806	54904.4363

Table 5: Bayesian deviance estimates for the three IRT models with the CBASE data.

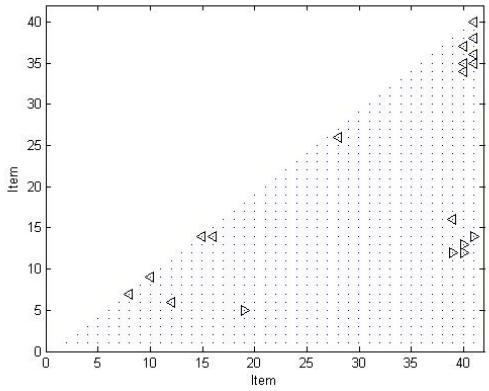
tation of the data. The three IRT models described in the previous section were then each fit to the CBASE data using Gibbs sampling with a run length of $10,000$ iterations and a burn-in period of $5,000$, which was sufficient for the chains to converge. The results with Bayesian deviance and PPMC, displayed in Table 5 and Figure 3, respectively, suggest that with a smaller Bayesian DIC value and fewer number of extreme PPP values for odds ratios, the multi-unidimensional model provided a relatively better description of the data, even after taking into consideration model complexity. In addition, the unidimesnional model had a clearly better fit to the data than the unidimensional model (2), indicating that the two latent traits should have a correlation greater than 0.5. Hence, the actual latent structure for the CBASE data is closer to multi-unidimensional with the two subtests being moderately to highly correlated, and a unified English trait is not sufficient in describing the specific trait levels necessary for the "writing" and "reading" subtests.

6. Discussion

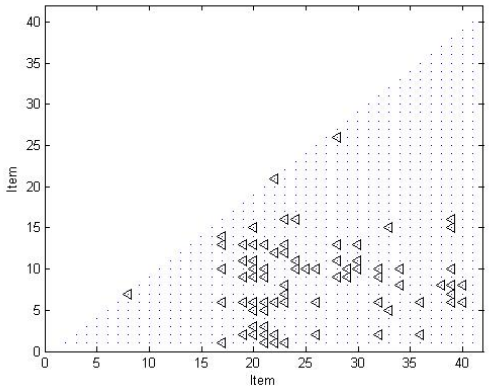
With functions for generating dichotomous response data from and implementing the Gibbs sampler for the 2PNO multi-unidimensional IRT model, **IRTmu2no** allows the user the choice to set the number of total or burn-in samples, specify starting values or prior distributions for model parameters, check convergence of the Markov chain, as well as obtain Bayesian model choice or model checking statistics. The package leaves it to the user to choose between noninformative and informative priors for the item parameters. In addition, the user can choose to set the location and scale parameters for the conjugate normal priors of α_{vj} and β_{vj} to reflect different prior beliefs on their distributions. For example, if there is a strong prior opinion that the item intercepts should be centered around 0, a smaller $\sigma^2_{\beta_v}$ can be specified in the `gsmu2no` function such that `gprior=[zeros(m,1),0.5*ones(m,1)]`. If, however, this prior opinion concerns with the item intercepts in the first subtest, not the entire test, the input argument becomes `gprior=[zeros(m,1),[0.5; ones(m-1,1)]]`. This way, different prior distributions can be specified for α_{vj} or β_{vj} across the m subtests.

One should note that during an implementation of the Gibbs sampler, if a Markov chain does not converge within a run length of certain iterations, additional iterations can be obtained by invoking the `gsmu2no` function with starting values `th0`, `item0`, and `sigma0` set to be their respective posterior samples drawn on the last iteration of the Markov chain (see Sheng 2008, for a demonstration of such procedure).

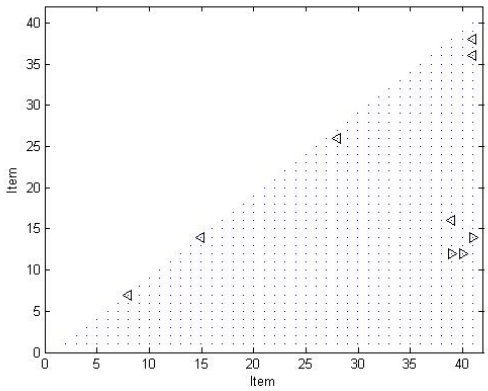
The illustrative examples provided in Section 5 deal with two subtests for simplicity. For tests with three or more subtests, **IRTmu2no** can be used in a similar fashion with possibly increased complexity and consequently a longer computing time. In addition, Bayesian deviance and



(a) unidimensional model



(b) unidimensional model (2)



(c) multi-unidimensional model

Figure 3: Plots of PPP values for odds ratios for the CBASE data fitted with the three IRT models. (Triangles represent extreme PPP values, where values smaller than 0.005 are denoted using left triangle signs and values larger than 0.995 are denoted using right triangle signs.)

PPMC are adopted in this paper to evaluate the goodness-of-fit of a candidate model. One may also want to consider Bayes factors, which provide more reliable and powerful results for model comparisons in the Bayesian framework. However, they are difficult to calculate due to the difficulty in exact analytic evaluation of the marginal density of the data (Kass and Raftery 1995) and hence are not considered in the paper. In addition, this paper adopts the Gelman-Rubin R statistic to assess convergence numerically. Its multivariate extension, the Brooks-Gelman multivariate potential scale reduction factor (Brooks and Gelman 1998), may be considered as well.

Acknowledgments

The author would like to thank the associate editor and two anonymous reviewers for their valuable comments and suggestions.

References

Adams RJ, Wilson M, Wang WC (1997). "The Multidimensional Random Coefficients Multinomial Logit Model." *Applied Psychological Measurement*, **21**, 1–23.

Albert JH (1992). "Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling." *Journal of Educational Statistics*, **17**, 251–269.

Bafumi J, Gelman A, Park DK, Kaplan N (2005). "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis Advance Access*, **13**, 171–187.

Bezruckzo N (ed.) (2005). *Rasch Measurement in Health Sciences*. JAM Press, Maple Grove, MN.

Birnbaum A (1968). "The Logistic Test Model." In FM Lord, MR Novick (eds.), "Statistical Theories of Mental Test Scores," pp. 397–423. Addison-Wesley.

Birnbaum A (1969). "Statistical Theory for Logistic Mental Test Models with a Prior Distribution of Ability." *Journal of Mathematical Psychology*, **6**, 258–276.

Bock RD, Aitkin M (1981). "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm." *Psychometrika*, **46**, 443–459.

Brooks SP, Gelman A (1998). "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Carlin BP, Louis TA (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd edition. Chapman & Hall, London.

Casella G, George EI (1992). "Explaining the Gibbs Sampler." *The American Statistician*, **46**, 167–174.

Chang CH, Reeve BB (2005). "Item Response Theory and Its Applications to Patient-Reported Outcomes Measurement." *Evaluation & the Health Professions*, **28**, 264–282.

Chib S, Greenberg E (1995). "Understanding the Metropolis-Hastings Algorithm." *The American Statistician*, **49**, 327–335.

Embretson SE, Reise SP (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, New Jersey.

Feske U, Kirisci L, Tarter RE, Plkonis PA (2007). "An Application of Item Response Theory to the DSM-III-R Criteria for Borderline Personality Disorder." *Journal of Personality Disorders*, **21**, 418–433.

Gelfand AE, Smith AFM (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, **85**, 398–409.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton.

Gelman A, Meng XL, Stern HS (1996). "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies." *Statistica Sinica*, **6**, 733–807.

Gelman A, Rubin DB (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science*, **7**, 457–511.

Geman S, Geman D (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer GJ (1992). "Practical Markov Chain Monte Carlo." *Statistical Science*, **7**, 473–483.

Imbens GW (2000). "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika*, **87**, 706–710.

Johnson VE, Albert JH (1999). *Ordinal Data Modeling*. Springer-Verlag, New York.

Kass RE, Raftery AE (1995). "Bayes Factors." *Journal of the American Statistical Association*, **90**, 773–795.

Kolen MJ, Brennan RL (eds.) (1995). *Test Equating: Methods and Practices*. Springer-Velag, New York.

Lee H (1995). *Markov Chain Monte Carlo Methods for Estimating Multidimensional Ability in Item Response Analysis*. Ph.D. thesis, University of Missouri, Columbia, MO.

Li YH, Schafer WD (2005). "Trait Parameter Recovery Using Multidimensional Computerized Adaptive Testing in Reading and Mathematics." *Applied Psychological Measurement*, **29**, 3–25.

Lord FM (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

McDonald RP (1999). *Test Theory: A Unified Approach*. Lawrence Erlbaum, Mahwah, NJ.

Molenaar IW (1995). "Estimation of Item Parameters." In GH Fischer, IW Molenaar (eds.), "Rasch Models: Foundations, Recent Developments, and Applications," pp. 39–51. Springer-Verlag, New York.

Osterlind SJ (1997). "A National Review of Scholastic Achievement in General Education: How Are We Doing and Why Should We Care?" *ASHE-ERIC Higher Education Report Volume 25, No. 8*, George Washington University Graduate School of Education and Human Development, Washington, DC.

Patz RJ, Junker BW (1999a). "Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses." *Journal of Educational and Behavioral Statistics*, **24**, 342–366.

Patz RJ, Junker BW (1999b). "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models." *Journal of Educational and Behavioral Statistics*, **24**, 146–178.

Reckase MD (1997). "The Past and Future of Multidimensional Item Response Theory." *Applied Psychological Measurement*, **21**, 25–36.

Ripley BD (1987). *Stochastic Simulation*. John Wiley & Sons, New York.

Rubin DB (1984). "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." *Annals of Statistics*, **12**, 1151–1172.

Sheng Y (2008). "Markov Chain Monte Carlo Estimation of Normal Ogive IRT Models in MATLAB." *Journal of Statistical Software*, **25**, 1–15. URL http://www.jstatsoft.org/v25/i08/.

Sheng Y, Wikle CK (2007). "Comparing Multiunidimensional and Unidimensional IRT Models." *Educational & Psychological Measurement*, **67**, 899–919.

Sinharay S, Johnson MS, Stern HS (2006). "Posterior Predictive Assessment of Item Response Theory Models." *Applied Psychological Measurement*, **30**, 298–321.

Sinharay S, Stern HS (2002). "On the Sensitivity of Bayes Factors to the Prior Distribution." *The American Statistician*, **56**, 196–201.

Sinharay S, Stern HS (2003). "Posterior Predictive Model Checking in Hierarchical Models." *Journal of Statistical Planning and Inference*, **111**, 209–221.

Spiegelhalter D, Best N, Carlin B (1998). "Bayesian Deviance, the Effective Number of Parameters, and the Comparison of Arbitrarily Complex Models." *Technical Report 98-009*, Division of Biostatistics, University of Minnesota.

Tanner MA, Wong WH (1987). "The Calculation of Posterior Distribution by Data Augmentation." *Journal of the American Statistical Association*, **82**, 528–550.

The MathWorks, Inc (2007). *MATLAB – The Language of Technical Computing, Version 7.5*. The MathWorks, Inc., Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/.

Verdinelli I, Wasserman L (1995). "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio." *Journal of the American Statistical Association*, **90**, 614–618.

Wainer H, Dorans N, Eignor D, Flaugher R, Green B, Mislevy R, Steinberg L, Thissen D (eds.) (2000). *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Affiliation:

Yanyan Sheng
Department of Educational Psychology & Special Education
Wham 223, Mail Code 4618
Southern Illinois University-Carbondale
Carbondale, IL 62901, United States of America
E-mail: ysheng@siu.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 28, Issue 10 *Submitted:* 2008-04-22
November 2008 *Accepted:* 2008-11-08