Journal of Statistical Software
April 2008, Volume 25, Issue 8. http://www.jstatsoft.org/

Markov Chain Monte Carlo Estimation of Normal Ogive IRT Models in MATLAB

Yanyan Sheng
Southern Illinois University-Carbondale

Abstract

Modeling the interaction between persons and items at the item level for binary response data, item response theory (IRT) models have been found useful in a wide variety of applications in various fields. This paper provides the requisite information and description of software that implements the Gibbs sampling procedures for the one-, two- and three-parameter normal ogive models. The software developed is written in the MATLAB package **IRTuno**. The package is flexible enough to allow a user the choice to simulate binary response data, set the number of total or burn-in iterations, specify starting values or prior distributions for model parameters, check convergence of the Markov chain, and obtain Bayesian fit statistics. Illustrative examples are provided to demonstrate and validate the use of the software package. The m-file `v25i08.m` is also provided as a guide for the user of the MCMC algorithms with the three dichotomous IRT models.

Keywords: item response theory, unidimensional normal orgive models, MCMC, Gibbs sampling, Gelman-Rubin R, Bayesian DIC, MATLAB.

1. Introduction

Item response theory (IRT) provides a collection of models that describe how items and persons interact to yield probabilistic correct/incorrect responses. The influence of items and persons on the responses is modeled by distinct sets of parameters so that the probability of a correct response to an item is a function of the person's latent trait, θ_i, and the item's characteristics, $\boldsymbol{\xi}_j$, i.e.,

$$P(y = \text{correct}) = f(\theta_i, \boldsymbol{\xi}_j).$$

The model assumes one θ_i parameter for each person and is commonly referred to as the unidimensional model, signifying that each test item measures some facet of the unified latent trait. Much research has been conducted on the development and application of such IRT

models in educational and psychological measurement (e.g., Bock and Aitkin 1981; Mislevy 1985; Patz and Junker 1999b; Tsutakawa and Lin 1986).

Studies in other fields have also utilized them or similar models in a wide variety of applications (e.g., Bafumi, Gelman, Park, and Kaplan 2005; Bezruckzo 2005; Chang and Reeve 2005; Feske, Kirisci, Tarter, and Plkonis 2007; Fienberg, Johnson, and Junker 1999; Imbens 2000; Reiser 1989; Sinharay and Stern 2002).

Simultaneous estimation of both item and person parameters in IRT models results in statistical complexities in the estimation task, as consistent estimates are not available. This problem has made estimation procedure a primary focus of psychometric research over decades (Birnbaum 1969; Bock and Aitkin 1981; Molenaar 1995). Recent attention is focused on Markov chain Monte Carlo (MCMC, see e.g., Chib and Greenberg 1995) simulation techniques, which have been influential in modern Bayesian analyses where they are used to summarize the posterior distributions that arise in the context of the Bayesian prior-posterior framework (Carlin and Louis 2000; Chib and Greenberg 1995; Gelfand and Smith 1990; Gelman, Carlin, Stern, and Rubin 2004; Tanner and Wong 1987). MCMC methods have proved useful in practically all aspects of Bayesian inference, such as parameter estimation and model comparisons. A key reason for the widespread interest in them is that they are extremely general and flexible and hence can be used to sample univariate and multivariate distributions when other methods (e.g., marginal maximum likelihood) either fail or are difficult to implement. In addition, MCMC allows one to model the dependencies among parameters and sources of uncertainty (Tsutakawa and Johnson 1990; Tsutakawa and Soltys 1988).

Albert (1992)—see also Baker (1998)—was the first to apply an MCMC algorithm, known as Gibbs sampling (Chib and Greenberg 1995; Gelfand and Smith 1990; Geman and Geman 1984), to the two-parameter normal ogive (2PNO, Lord and Novick 1968) IRT model using the data augmentation idea of Tanner and Wong (1987). Johnson and Albert (1999) further generalized the approach to the three-parameter normal ogive (2PNO, Lord 1980) IRT model. With no prior information concerning the item parameters, noninformative priors were adopted in their implementation so that inference was based solely on the data. However, in some applications, informative priors are more preferred than noninformative priors. This is especially the case with the 3PNO model, where improper noninformative priors are to be avoided given the reason described later in this paper. Moreover, when comparing several candidate models, Bayes factors are commonly adopted in the Bayesian framework, but they are not defined with noninformative priors (Gelman *et al.* 2004). Studies have also shown that by incorporating prior information about item parameters, model parameters can be estimated more accurately with smaller sample sizes (Mislevy 1986; Swaminathan and Gifford 1983, 1985, 1986).

In view of the above, the objective of this paper is to provide a MATLAB (The MathWorks, Inc. 2007) package that implements Gibbs sampling procedures for the one-, two-, and three-parameter normal ogive IRT models with the option of specifying noninformative or informative priors for item parameters. Section 2 reviews the three unidimensional models. Section 3 briefly describes the MCMC algorithms which are implemented in the package **IRTuno**. In Section 4, a brief illustration is given of a Bayesian model selection technique for testing the fit of the model. The package **IRTuno** is introduced in Section 5, where a description is given of common input and output variables. In Section 6, illustrative examples are provided to demonstrate the use of the source code. Finally, a few summary remarks are given in Section 7.

It has to be noted that more complicated MCMC procedures have to be adopted for the logistic form of IRT models. For example, Patz and Junker (1999a,b) adopted the Metropolis-Hastings within Gibbs (Chib and Greenberg 1995) for the two-parameter logistic (2PL) and the three-parameter logistic (3PL) models. 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), the MCMC procedures for logistic models are not considered in this paper.

2. IRT models

The unidimensional IRT model provides a fundamental framework in modeling the person-item interaction by assuming one latent trait. Suppose a test consists of k dichotomous (0-1) items, each measuring a single unified trait, θ. Let $\mathbf{y} = [y_{ij}]_{n\times k}$ represent a matrix of n examinees' responses to the k items, so that y_{ij} is defined as

$$y_{ij} = \begin{cases} 1, & \text{if person } i \text{ answers item } j \text{ correctly} \\ 0, & \text{if person } i \text{ answers item } j \text{ incorrectly} \end{cases}$$

for $i = 1, ..., n$ and $j = 1, .., k$.

The probability of person i obtaining correct response for item j can be defined as

$$P(y_{ij} = 1|\theta_i, \beta_j) = \Phi(\theta_i - \beta_j) = \int_{-\infty}^{\theta_i-\beta_j} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \tag{1}$$

for the one-parameter normal ogive (1PNO) model, where β_j is associated with item difficulty, θ_i is a scalar latent trait parameter, and the term "one-parameter" indicates that there is one item parameter β_j in the model;

$$P(y_{ij} = 1|\theta_i, \alpha_j, \beta_j) = \Phi(\alpha_j\theta_i - \beta_j) \tag{2}$$

for the 2PNO model, where α_j is a positive scalar parameter describing the item discrimination; and

$$\begin{aligned} P(y_{ij} = 1|\theta_i, \alpha_j, \beta_j, \gamma_j) &= \gamma_j + (1 - \gamma_j)\Phi(\alpha_j\theta_i - \beta_j) \\ &= \Phi(\alpha_j\theta_i - \beta_j) + \gamma_j(1 - \Phi(\alpha_j\theta_i - \beta_j)), \quad 0 \le \gamma_j < 1 \end{aligned} \tag{3}$$

for the 3PNO model, where γ_j is a pseudo-chance-level parameter, indicating that the probability of correct response is greater than zero even for those with very low trait levels. The 3PNO model is applicable for objective items such as multiple-choice or true-or-false items where an item is too difficult for some examinees. As one may note, the three models are increasingly more general and hence more complex.

3. MCMC algorithms

Gibbs sampling is one of the simplest MCMC algorithms that are used to obtain item and person parameter estimates simultaneously. 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. Its general underlying strategy is to iteratively sample each item parameter, e.g., $\boldsymbol{\xi}_j$, where $\boldsymbol{\xi}_j = (\alpha_j, \beta_j)'$, and person parameter, θ_i, from their respective posterior distributions, conditional on the sampled values of all other person and item parameters, with starting values $\boldsymbol{\xi}_j^{(0)}$ and $\theta_i^{(0)}$. This iterative process continues for a sufficient number of samples after the posterior distributions converge to stationary distributions (a phase known as burn-in).

As with standard Monte Carlo, the posterior means of all the samples collected after the burn-in stage are considered as estimates of the true parameters $\boldsymbol{\xi}_j$ and θ_i. 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).

The Gibbs sampler for each IRT model is now briefly described. For ease of illustration, Section 3.1 starts with the more general 3PNO model.

3.1. Gibbs sampling for the 3PNO model

To implement Gibbs sampling for the 3PNO model defined in (3), two latent variables, Z and W, are introduced such that $Z_{ij} \sim N(\eta_{ij}, 1)$ (Albert 1992; Tanner and Wong 1987), where $\eta_{ij} = \alpha_j \theta_i - \beta_j$, and

$$W_{ij} = \begin{cases} 1, & \text{if person } i \text{ knows the correct answer to item } j \\ 0, & \text{if person } i \text{ doesn't know the correct answer to item } j \end{cases}$$

with a probability density function

$$P(W_{ij} = w_{ij}|\eta_{ij}) = \Phi(\eta_{ij})^{w_{ij}}(1 - \Phi(\eta_{ij}))^{1-w_{ij}}. \tag{4}$$

With prior distributions assumed for θ_i , $\boldsymbol{\xi}_j$ and γ_j, the joint posterior distribution of ($\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\boldsymbol{\gamma}$, $\mathbf{W}$, $\mathbf{Z}$) is hence

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \mathbf{W}, \mathbf{Z}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{W}, \boldsymbol{\gamma}) p(\mathbf{W}|\mathbf{Z}) p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\theta}) p(\boldsymbol{\xi}) p(\boldsymbol{\gamma}), \tag{5}$$

where

$$f(\mathbf{y}|\mathbf{W},\boldsymbol{\gamma}) = \prod_{i=1}^{n}\prod_{j=1}^{k} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \tag{6}$$

is the likelihood function, with p_{ij} being the probability function for the 3PNO model as defined in (3).

In the literature, noninformative priors for the item parameters have been adopted in the description of the model specification (see e.g., Béguin and Glas 2001; Glas and Meijer 2003). However, they result in an improper posterior distribution because the 3PNO model, as defined in (3), can be viewed as a mixture model with two components, i.e., 1 and $\Phi(\alpha_j\theta_i-\beta_j)$, so the probability of an observation coming from a component is γ_j or $1-\gamma_j$. Although the prior for the non-component parameter of the mixture (γ_j in this context) can be chosen in a typical fashion (Richardson and Green 1997), improper noninformative priors for component specific parameters are not recommended, as the joint posterior distribution is not defined and parameter estimates are likely to be unstable (Diebolt and Robert 1994). Hence, with a normal prior for θ_i , a conjugate Beta prior for γ_j and informative conjugate priors for α_j and β_j so that $\theta_i \sim N(\mu,\sigma^2)$, $\gamma_j \sim Beta(s_j,t_j)$, $\alpha_j \sim N_{(0,\infty)}(\mu_\alpha,\sigma_\alpha^2)$ and $\beta_j \sim N(\mu_\beta,\sigma_\beta^2)$, the full conditional distributions of W_{ij}, Z_{ij}, θ_i, $\boldsymbol{\xi}_j$ and γ_j can be derived in closed forms as follows:

$$W_{ij}|\bullet \sim \begin{cases} Bernoulli\left(\frac{\Phi(\eta_{ij})}{\gamma_j+(1-\gamma_j)\Phi(\eta_{ij})}\right), & \text{if} \quad y_{ij}=1 \\ Bernoulli(0), & \text{if} \quad y_{ij}=0 \end{cases}, \tag{7}$$

$$Z_{ij}|\bullet \sim \begin{cases} N_{(0,\infty)}(\eta_{ij},1), & \text{if} \quad W_{ij}=1 \\ N_{(-\infty,0)}(\eta_{ij},1), & \text{if} \quad W_{ij}=0 \end{cases}, \tag{8}$$

$$\theta_i|\bullet \sim N\left(\frac{\sum_j (Z_{ij}+\beta_j)\alpha_j+\mu/\sigma^2}{1/\sigma^2+\sum_j \alpha_j^2}, \frac{1}{1/\sigma^2+\sum_j \alpha_j^2}\right), \tag{9}$$

$$\boldsymbol{\xi}_j|\bullet \sim N((\mathbf{x}'\mathbf{x}+\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1})^{-1}(\mathbf{x}'\mathbf{Z}_j+\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\xi}}),(\mathbf{x}'\mathbf{x}+\boldsymbol{\Sigma}_{\boldsymbol{\xi}}^{-1})^{-1})I(\alpha_j>0), \tag{10}$$

where $\mathbf{x}=[\boldsymbol{\theta},-1]$, $\boldsymbol{\mu}_{\boldsymbol{\xi}}=(\mu_\alpha,\mu_\beta)'$, and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\beta^2 \end{pmatrix}$,

$$\gamma_j|\bullet \sim Beta(a_j+s_j, b_j-a_j+t_j), \tag{11}$$

where s_j and t_j are the two parameters in the specified Beta prior distribution, a_j denotes the number of persons who do not know the correct answer to item j, and b_j denotes the number of correct responses obtained by guessing. It has to be noted that when $s_j=t_j=1$, the prior distribution for γ_j is uniform. Higher values of these parameters lead to more precise prior information. In practice, these parameters should be chosen in a way that $E(\gamma_j)=\frac{s_j}{s_j+t_j}$ is some pre-specified value (see e.g., Swaminathan and Gifford 1986). It is also noted that the proper normal prior for θ_i with specific μ and σ^2 values (e.g., $\mu=0$ and $\sigma^2=1$) ensures unique scaling and hence is essential in resolving a particular identification problem in the model (see e.g., Albert 1992, for a description of the problem).

Hence, with starting values $\theta^{(0)}$, $\boldsymbol{\xi}^{(0)}$ and $\gamma^{(0)}$, observations ($W^{(l)}$, $Z^{(l)}$, $\theta^{(l)}$, $\boldsymbol{\xi}^{(l)}$, $\gamma^{(l)}$) can be simulated using the Gibbs sampler by iteratively drawing from their respective full conditional distributions specified in (7) through (11).

3.2. Gibbs sampling for the 2PNO model

Assuming that no guessing is involved in item responses (i.e., $\gamma_j = 0$), the 2PNO model, without being a mixture model, is much simpler than the 3PNO model and hence the prior distributions can be chosen in a typical fashion. The Gibbs sampler involves updating samples for fewer model parameters, namely, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, and the augmented continuous variable $\mathbf{Z}$, where Z_{ij} is defined so that $y_{ij} = \begin{cases} 1, & \text{if} \quad Z_{ij} > 0 \\ 0, & \text{if} \quad Z_{ij} \leq 0 \end{cases}$. Their joint posterior distribution is

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Z}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{Z}) p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\theta}) p(\boldsymbol{\xi}). \tag{12}$$

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

$$Z_{ij}|\bullet \sim \begin{cases} N_{(0,\infty)}(\eta_{ij}, 1), & \text{if} \quad y_{ij} = 1 \\ N_{(-\infty,0)}(\eta_{ij}, 1), & \text{if} \quad y_{ij} = 0 \end{cases}, \tag{13}$$

a sampling of person trait θ from (9), and a sampling of the item parameters $\boldsymbol{\xi}$ from

$$\boldsymbol{\xi}_j|\bullet \sim N((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Z}_j, (\mathbf{x}'\mathbf{x})^{-1}) I(\alpha_j > 0) \tag{14}$$

assuming noninformative uniform priors $\alpha_j > 0$ and $p(\beta_j) \propto 1$, or from (10) assuming informative priors. See Albert (1992) for a detailed illustration of the procedure.

3.3. Gibbs sampling for the 1PNO model

Likewise, the MCMC algorithm for the 1PNO model, a special case of the 2PNO model, involves only $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\mathbf{Z}$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)'$. Their joint posterior distribution is

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{Z}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{Z}) p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\beta}) p(\boldsymbol{\theta}) p(\boldsymbol{\beta}). \tag{15}$$

The implementation of the Gibbs sampling procedure again involves three sampling processes: a sampling of the augmented Z parameters from (13), a sampling of person trait θ from (9), and a sampling of the item difficulty parameters β from $N(n^{-1}\sum_i (\theta_i - Z_{ij}), n^{-1})$ assuming a uniform prior $p(\beta_j) \propto 1$, or from $N\left((n + 1/\sigma_\beta^2)^{-1}[\mu_\beta/\sigma_\beta^2 + \sum_i (\theta_i - Z_{ij})], (n + 1/\sigma_\beta^2)^{-1}\right)$ assuming a conjugate normal prior $\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$.

4. Bayesian model choice technique

In the Bayesian framework, the adequacy of the fit of a given model is evaluated using several model choice techniques, among which, Bayesian deviance is considered and briefly illustrated. It should be noted that this measure provides a model comparison criterion. Hence, it evaluates the fit of a model in a relative, not absolute, sense.

The Bayesian deviance information criterion (DIC) was introduced by Spiegelhalter, Best, and Carlin (1998) who generalized the classical information criteria to one that is based on the posterior distribution of the deviance. This criterion is defined as DIC $= \overline{D} + p_D$, where $\overline{D} \equiv E_{\vartheta|y}(D) = E(-2\log L(y|\vartheta))$ is the posterior expectation of the deviance (with L

being the likelihood function, where ϑ denotes all model parameters) and $p_D = E_{\vartheta|y}(D) - D(E_{\vartheta|y}(\vartheta)) = \overline{D} - D(\bar{\vartheta})$ is the effective number of parameters (Carlin and Louis 2000). Further, $D(\bar{\vartheta}) = -2\log(L(y|\bar{\vartheta}))$, where $\bar{\vartheta}$ is the posterior mean. To compute Bayesian DIC, MCMC samples of the parameters, $\vartheta^{(1)}, \ldots, \vartheta^{(G)}$, can be drawn with the Gibbs sampling procedure, then $\overline{D}$ could be approximated as $\overline{D} = \frac{1}{G}(-2\log\prod_{g=1}^{G} L(y|\vartheta^{(g)}))$. 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. Indeed, this measure has been found useful for model comparisons in the IRT literature (e.g., Sheng and Wikle 2007, 2008).

5. Package IRTuno

The package **IRTuno** contains two major user-callable routines: a function for generating unidimensional IRT data using the 1PNO, 2PNO or 3PNO model titled, `simIRTuno`, and a function that implements MCMC to obtain posterior samples, estimates, convergence statistics, or model fit statistics, `gsIRTuno`. Both functions require the user to have access to the MATLAB Statistics Toolbox.

The function `simIRTuno` has input arguments `N`, `K`, `iparm`, and `p` for the number of respondents, the number of items, user-specified item parameters, and the IRT model based on which the binary response data is generated (`p=1` for the 1PNO model, `p=2` for the 2PNO model, and `p=3` for the 3PNO model), respectively. The optional `iparm` argument allows the user the choice to input the respective item parameters for each model, or randomly generate them from uniform distributions so that $\alpha_j \sim U(0, 2)$, $\beta_j \sim U(-2, 2)$, and/or $\gamma_j \sim U(.05, .6)$. The user can further choose to store the simulated person (`theta`) and item (`item`) parameters.

The function `gsIRTuno` initially reads in the starting values for the parameters (`th0`, `item0`), or sets them to be $\theta_i^{(0)} = 0$, $\alpha_j^{(0)} = 2$, $\beta_j^{(0)} = -\Phi^{-1}(\sum_i y_{ij}/n)\sqrt{5}$ (Albert 1992) for the 1PNO and/or 2PNO model, and $\theta_i^{(0)} = 0$, $\alpha_j^{(0)} = 1$, $\beta_j^{(0)} = 0$, $\gamma_j^{(0)} = 0.2$ (Béguin and Glas 2001) for the 3PNO model. It then implements the Gibbs sampler to a user-specified IRT model (`parm`) and iteratively draws random samples for the parameters from their respective full conditional distributions. The normal prior distribution for θ_i (`tprior`) can assume $\mu = 0$, $\sigma^2 = 1$ (default) or any values of interest. The prior distributions for the item parameters in the 1PNO and 2PNO models can be noninformative (`flat=1`) or informative (`flat=0`, default). In the latter case, the user can specify any values of interest or use the default values, namely, $\mu_\alpha = 0$ and $\sigma_\alpha^2 = 1$ for α_j (`aprior`), and $\mu_\beta = 0$ and $\sigma_\beta^2 = 1$ for β_j (`gprior`). Given the reason provided in Section 3.1, only informative priors can be specified for $\boldsymbol{\xi}_j$ (`flat=0`) in the 3PNO model, where the two parameters in the Beta prior distribution for γ_j (`cprior`) can be $s_j = 5$, $t_j = 7$ (default) or any values of interest. 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 and item estimates (`iparm`, `pparm`). 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 (`R`) with `gsIRTuno`. Moreover, 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. The function's input and output arguments are completely specified in the m-file.

6. Illustrative examples

To demonstrate the use of the **IRTuno** package, simulated and real data examples are provided in this section to illustrate item parameter recovery as well as model goodness of fit.

6.1. Parameter recovery

For parameter recovery, three 1000-by-10 (i.e., $n = 1000$ and $k = 10$) dichotomous data matrices were simulated from the 1PNO, 2PNO, and 3PNO models, respectively, using the item parameters from Béguin and Glas (2001, p. 551), which are shown in the first column of Tables 1, 2 and 3. The Gibbs sampler was implemented to recover the item parameters for the respective 1PNO and 2PNO models assuming the informative or noninformative prior distributions described previously, and those for the 3PNO model assuming informative priors. The posterior means and standard deviations for item parameters α, β, and/or γ, as well as their Monte Carlo standard errors of estimates and Gelman-Rubin R statistics were obtained and are displayed in the rest of the tables.

It has to be pointed out again that improper priors for component specific parameters in mixture models tend to result in unstable parameter estimates as the joint posterior distribution is not defined, and hence they are to be avoided for the 3PNO model. On the other hand, as informative priors lead to a proper posterior distribution for the model, valid estimates of item and person parameters can be obtained with sufficient iterations. Indeed, as Table 1 suggests, the Markov chain did reach stationarity with a run length of $30,000$ iterations and a burn-in period of $25,000$ iterations. Moreover, the posterior estimates of $\boldsymbol{\xi}_j$ are fairly close to the true parameters.

For the 1PNO and 2PNO models, stationarity was reached within $10,000$ iterations, and the item parameters were estimated with enough accuracy (see Tables 2 and 3). 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 α and/or β.

6.2. Model goodness of fit

To illustrate the use of the Bayesian deviance measure for the fit of a model, a 500-by-30 dichotomous data matrix was simulated from the 2PNO model using item parameters randomly drawn from the uniform distributions described in Section 5. The Gibbs sampler was subsequently implemented for each of the 1PNO, 2PNO and 3PNO models so that $10,000$ samples were simulated with the first $5,000$ set to be burn-in. It has to be noted that noninformative priors were adopted for the item parameters in the 1PNO and 2PNO models,

True	Estim.	SD	MCSE	R
α				
0.6400	0.5274	0.1107	0.0101	1.0227
0.8100	0.9135	0.1601	0.0139	1.0500
0.9600	1.0151	0.1593	0.0114	1.0348
0.9700	0.8363	0.1261	0.0132	1.0828
1.0500	0.8861	0.1655	0.0219	1.0258
0.8300	0.8911	0.1300	0.0168	1.1590
0.6100	0.7302	0.2022	0.0369	1.1968
0.8000	0.8510	0.1480	0.0142	1.0081
1.1700	1.6401	0.3540	0.0636	1.2070
1.5100	1.3186	0.3355	0.0837	1.3208
β				
−1.6200	−1.5409	0.1228	0.0143	1.0466
−1.5300	−1.4132	0.1230	0.0135	1.0221
−1.2900	−1.1354	0.1393	0.0150	1.0897
−1.0600	−0.9190	0.1055	0.0048	1.0157
−0.2400	−0.3587	0.1593	0.0225	1.0195
−0.2600	−0.2100	0.1159	0.0135	1.1426
0.0200	0.3587	0.2430	0.0483	1.2335
0.2100	0.3725	0.1569	0.0170	1.0273
−0.6700	−0.5156	0.1508	0.0139	1.0213
0.4800	0.3089	0.2322	0.0523	1.2459
γ				
0.1900	0.2473	0.0938	0.0070	1.0224
0.1500	0.2508	0.0886	0.0084	1.0294
0.1100	0.2592	0.0933	0.0097	1.0812
0.1400	0.2168	0.0775	0.0053	1.0412
0.3700	0.2829	0.0861	0.0123	1.0240
0.1400	0.1452	0.0548	0.0069	1.1692
0.1700	0.2876	0.0774	0.0149	1.2405
0.1000	0.1495	0.0477	0.0050	1.0270
0.1900	0.3225	0.0623	0.0086	1.0845
0.3100	0.2330	0.0567	0.0119	1.2095

Table 1: Posterior estimate, standard deviation (SD), Monte Carlo standard error of the estimate (MCSE) and Gelman-Rubin R statistic for each item parameter (α_j, β_j, γ_j) in the 3PNO model assuming informative priors (chain length = 30,000, burn-in = 25,000).

whereas informative priors were adopted for the 3PNO model. 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 both the posterior deviance (column 2) and the Bayesian DIC (column 5) estimates pointed to

	Informative priors				Noninformative priors			
True	Estim.	SD	MCSE	R	Estim.	SD	MCSE	R
α								
0.6400	0.5347	0.0835	0.0038	1.011	0.5525	0.0874	0.0039	1.006
0.8100	0.7401	0.0920	0.0069	1.027	0.7732	0.0955	0.0089	1.080
0.9600	1.0416	0.1058	0.0075	1.042	1.0562	0.1071	0.0069	1.009
0.9700	0.8507	0.0847	0.0053	1.009	0.8623	0.0839	0.0058	1.003
1.0500	1.1051	0.0907	0.0083	1.044	1.1257	0.0982	0.0067	1.035
0.8300	0.8065	0.0706	0.0019	1.000	0.8186	0.0719	0.0032	1.014
0.6100	0.5217	0.0557	0.0021	1.005	0.5287	0.0568	0.0012	1.000
0.8000	0.7460	0.0682	0.0030	1.000	0.7538	0.0656	0.0030	1.014
1.1700	1.1216	0.0979	0.0075	1.010	1.1486	0.0979	0.0079	1.065
1.5100	1.8401	0.2254	0.0156	1.031	1.8792	0.2197	0.0294	1.088
β								
−1.6200	−1.6178	0.0831	0.0023	1.004	−1.6396	0.0847	0.0047	1.014
−1.5300	−1.5257	0.0862	0.0051	1.023	−1.5591	0.0891	0.0083	1.078
−1.2900	−1.4498	0.0967	0.0079	1.050	−1.4647	0.0971	0.0055	1.008
−1.0600	−1.0166	0.0661	0.0040	1.003	−1.0305	0.0731	0.0045	1.003
−0.2400	−0.2165	0.0590	0.0028	1.004	−0.2281	0.0585	0.0026	1.009
−0.2600	−0.3553	0.0522	0.0011	1.001	−0.3628	0.0537	0.0018	1.005
0.0200	0.0688	0.0445	0.0009	1.002	0.0641	0.0457	0.0015	1.006
0.2100	0.1679	0.0490	0.0015	1.002	0.1608	0.0502	0.0013	1.002
−0.6700	−0.7231	0.0677	0.0031	1.009	−0.7438	0.0684	0.0046	1.024
0.4800	0.6369	0.1048	0.0060	1.012	0.6231	0.0950	0.0095	1.015

Table 2: Posterior estimate, standard deviation (SD), Monte Carlo standard error of the estimate (MCSE) and Gelman-Rubin R statistic for each item parameter (α_j, β_j) in the 2PNO model assuming informative and noninformative priors (chain length = 10,000, burn-in = 5,000).

the correct 2PNO model, indicating that it provided the best fit among the three candidate models.

6.3. 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 checking model goodness of fit. The data contains binary responses of 600 independent college students to a total of 41 English multiple-choice items. It is noted that in real test situations, the actual structure is not readily known. Hence, model comparison is necessary for establishing the model that provides a relatively better representation of the data. The 1PNO, 2PNO and 3PNO models were 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 Bayesian deviance estimates for each model, shown in Table 5, suggest that the 3PNO model provided a slightly better description of the data, even after taking into consideration model complexity.

	Informative priors				Noninformative priors			
True β	Estim.	SD	MCSE	R	Estim.	SD	MCSE	R
−1.6200	−1.5697	0.0671	0.0027	1.007	−1.5889	0.0679	0.0049	1.016
−1.5300	−1.5124	0.0671	0.0012	0.999	−1.5254	0.0673	0.0032	1.004
−1.2900	−1.3223	0.0651	0.0035	1.010	−1.3368	0.0633	0.0025	1.007
−1.0600	−1.0644	0.0603	0.0021	1.006	−1.0752	0.0607	0.0029	1.000
−0.2400	−0.2727	0.0557	0.0017	1.000	−0.2845	0.0559	0.0033	1.005
−0.2600	−0.2282	0.0547	0.0017	0.999	−0.2379	0.0567	0.0023	1.005
0.0200	0.1046	0.0559	0.0019	1.004	0.0979	0.0567	0.0024	1.001
0.2100	0.2222	0.0546	0.0024	1.003	0.2155	0.0565	0.0021	0.999
−0.6700	−0.6930	0.0562	0.0020	1.003	−0.7026	0.0556	0.0033	1.015
0.4800	0.5257	0.056	0.0023	1.002	0.5181	0.0582	0.0026	1.000

Table 3: Posterior estimate, standard deviation (SD), Monte Carlo standard error of the estimate (MCSE) and Gelman-Rubin R statistic for each item parameter (β_j) in the 1PNO model assuming informative and noninformative priors (chain length = 10,000, burn-in = 5,000)

	$\overline{D}$	$D(\bar{\vartheta})$	p_D	DIC
1PNO model	11276.1118	10790.8857	485.2261	11761.3379
2PNO model	10652.4105	10131.7251	520.6854	11173.0959
3PNO model	10840.1893	10319.3541	520.8352	11361.0245

Table 4: Bayesian deviance estimates for the 1PNO, 2PNO and 3PNO models with the simulated data.

	$\overline{D}$	$D(\bar{\vartheta})$	p_D	DIC
1PNO model	27065.2274	26452.6542	612.5733	27677.8007
2PNO model	26930.4079	26360.0000	569.9194	27500.3273
3PNO model	26874.7593	26281.9512	592.8081	27467.5674

Table 5: Bayesian deviance estimates for the 1PNO, 2PNO and 3PNO models with the CBASE data.

7. Discussion

With functions for generating dichotomous response data and implementing the Gibbs sampler for a user-specified unidimensional IRT model, **IRTuno** allows the user the choice to set the numbers of total and burn-in samples, specify starting values and prior distributions for model

parameters, check convergence of the Markov chain, as well as obtain Bayesian fit statistics. The package leaves it to the user to choose between noninformative and informative priors for the item parameters in the 1PNO and 2PNO models. In addition, the user can choose to set the location and scale parameters for the prior normal distributions of θ_i, α_j and β_j, or the two parameters in the Beta prior for γ_j to reflect different prior beliefs on their distributions. For example, if there is a stronger prior opinion that the item difficulties should be centered around 0, a smaller σ_β^2 can be specified in the `gsIRTuno` function such that `gprior=[0,0.5]`.

It is noted 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 `gsIRTuno` function with starting values `th0` and `item0` set to be their respective posterior samples drawn on the last iteration of the Markov chain. This is demonstrated in the example of assessing parameter recovery for the 3PNO model in `v25i08.m`.

In the CBASE example, one may note the small difference between the two DIC estimates for the 2PNO and 3PNO models from Table 5. Given the complexity of the 3PNO model, it is, however, not clear if this small difference makes a practical significance. Consequently, 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 for assessing convergence. 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 two anonymous reviewers for their valuable comments and suggestions.

References

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.

Baker FB (1998). "An Investigation of the Item Parameter Recovery Characteristics of a Gibbs Sampling Procedure." *Applied Psychological Measurement*, **22**, 163–169.

Béguin AA, Glas CAW (2001). "MCMC Estimation and Some Model-fit Analysis of Multidimensional IRT Models." *Psychometrika*, **66**, 541–562.

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

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

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*. Chapman & Hall, London, 2nd edition.

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.

Diebolt J, Robert CP (1994). "Estimation of Finite Mixture Distributions through Bayesian Sampling." *Journal of the Royal Statistical Society B*, **56**, 363–375.

Embretson SE, Reise SP (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ.

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.

Fienberg SE, Johnson MS, Junker BW (1999). "Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists." *Journal of the Royal Statistical Society A*, **162**, 383–392.

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, FL.

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

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

Glas CW, Meijer RR (2003). "A Bayesian Approach to Person Fit Analysis in Item Response Theory Models." *Applied Psychological Measurement*, **27**, 217–233.

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.

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

Lord FM, Novick MR (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

Mislevy RJ (1985). "Estimation of Latent Group Effects." *Journal of the American Statistical Association*, **80**, 993–997.

Mislevy RJ (1986). "Bayes Modal Estimation in Item Response Models." *Psychometrika*, **51**, 177–195.

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 S (1997). *A National Review of Scholastic Achievement in General Education: How are We Doing and Why Should We Care?*, volume 25 of *ASHE-ERIC Higher Education Report*. 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.

Reiser M (1989). "An Application of the Item-Response Model to Psychiatric Epidemiology." *Sociological Methods & Research*, **18**, 66–103.

Richardson S, Green PJ (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components." *Journal of the Royal Statistical Society B*, **59**, 731–792.

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

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

Sheng Y, Wikle CK (2008). "Bayesian Multidimensional IRT Models with a Hierarchical Structure." *Educational & Psychological Measurement*. In press, URL http://epm.sagepub.com/cgi/content/abstract/0013164407308512v1.

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

Spiegelhalter DJ, Best N, Carlin BP (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.

Swaminathan H, Gifford JA (1983). "Estimation of Parameters in the Three-parameter Latent Trait Model." In D Weiss (ed.), "New Horizons in Testing," pp. 13–30. Academic Press, New York.

Swaminathan H, Gifford JA (1985). "Bayesian Estimation in the Two-parameter Logistic Model." *Psychometrika*, **50**, 175–191.

Swaminathan H, Gifford JA (1986). "Bayesian Estimation in the Three-parameter Logistic Model." *Psychometrika*, **51**, 581–601.

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

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

Tsutakawa RK, Johnson JC (1990). "The Effect of Uncertainty of Item Parameter Estimation on Ability Estimates." *Psychometrika*, **55**, 371–390.

Tsutakawa RK, Lin HY (1986). "Bayesian Estimation of Item Response Curves." *Psychometrika*, **51**, 251–267.

Tsutakawa RK, Soltys MJ (1988). "Approximation for Bayesian Ability Estimation." *Journal of Educational Statistics*, **13**, 117–130.

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.

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 25, Issue 8 *Submitted:* 2007-09-05
April 2008 *Accepted:* 2008-03-27