A Review of BMIRT TOOLKIT: BMIRT for Bayesian Multivariate IRT

Applied Psychological Measurement

2015, Vol. 39(2) 155–159

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621614542607

apm.sagepub.com

Richard D. Schwarz¹

Abstract

A software review was conducted for BMIRT (Bayesian Multivariate IRT) that implements a variety of multidimensional models using Markov chain Monte Carlo (MCMC) methods. The review describes its basic functionality, implementation of the program, and the operating environment.

Keywords

dimensionality, MCMC, MIRT

Improvements in analytic techniques and computational resources currently permit the estimation of models with a high number of parameters that can be applied to large, complex test designs. Multidimensional models have been proposed in the measurement literature for decades, but until recently, there have not been truly general-purpose programs that can model a variety of complex data types consisting of dichotomous and polytomous items or dependent structures such as reading passages. The Bayesian Multivariate Item Response Theory (**BMIRT**) **TOOLKIT** is a multipurpose program that consists of a suite of procedures that provide researchers with methods typically used to address a variety of solutions associated with multidimensional models. It consists of three associated software programs: a MIRT calibration program (**BMIRT II**), a linking program (**LinkMIRT**), and simulation programs for multidimensional fixed forms and computer adaptive testing (**SimuMIRT**; **SimuMCAT**). For this review, only the core program **BMIRT II** is evaluated.

General Program Functionality

Within the IRT model framework, **BMIRT II** (i.e., **BMIRT**) is a multipurpose program that supports both unidimensional or multidimensional versions of the Rasch, the two- and three-parameter logistic, generalized two-parameter partial-credit, graded response, higher order linear models, multidimensional testlet, and rater effects models. Noncompensatory MIRT models are also available. Using full Bayesian estimation, both item parameters and ability estimates are produced simultaneously. Both exploratory (complex structure) and confirmatory modes (simple structure) or a combination are supported. The assignment of items to different factor

¹Educational Testing Service, San Francisco, CA, USA

Corresponding Author:

Richard D. Schwarz, Educational Testing Service, 90 New Montgomery Street (Suite 1500), San Francisco, CA 94105, USA.

Email: RSchwarz@ets.org

models is implemented using a simple design matrix. Factor loadings are indicated on a per item basis by either zero or one in a design matrix in the program control file. Each row of the design matrix corresponds to a factor. For example, a bifactor model is specified by indicating all items loading on the major factor with subsets of the items assigned to the minor factors. Single-group or multigroup calibration approaches consistent with vertical scaling applications can be implemented. The structure of data files needed for **BMIRT** use the familiar examinee-by-item response arrays typically found with many types of scaling programs. From a practical perspective, an important requirement for any IRT calibration program is the capability to analyze large, sparse response data matrices (e.g., for assessment programs that employ many test forms). **BMIRT** makes a provision for this type of missing data with a “not-reached” code. Using the item parameters, three types of overall multidimensional ability estimates are computed: (a) maximum a posteriori estimates (MAP), (b) expected a posteriori (EAP), and (c) maximum likelihood estimates (MLE; Yao, 2014). By fixing a set of item parameters, other groups of examinees can be scored—for example, in a pre-equating context where the calibrations are performed at an earlier time because of tight score reporting timelines. It provides associated statistics such as multidimensional item and test information functions and classification, accuracy, and consistency indices (Grima & Yao, 2011). Differential item functioning (DIF) statistics applied within a multidimensional framework are also available (Yao & Li, 2010).

In exploratory factor analysis, an important step is to determine the number and types of factors to be retained. Candidate models are proposed with different structures where global fit is evaluated to arrive at a preferred version. **BMIRT** provides a variety of well-known global fit statistics such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and deviance information criterion (DIC) that impose a penalty for adding extra parameters. Cluster analysis of angle measures (Luecht & Miller, 1992; Miller & Hirsch, 1992; Reckase, 2009) is provided by the program as an alternative method for assessing dimensional structure. Hypotheses concerning the number of factors to retain and how construct shifts over levels of the test can be examined using a multigroup approach. The program permits a multigroup calibration approach consistent with vertical scaling to be implemented within the **MIRT** framework.

Model Estimation

Estimation can be problematic with multidimensional models due to the inclusion of many additional parameters compared with the unidimensional case. To estimate complex models of that are highly parameterized, **BMIRT** uses the Markov chain Monte Carlo (MCMC) Metropolis–Hastings method to produce item and ability parameter estimates simultaneously. Metropolis–Hastings is a general term for Markov chain simulation methods, which are useful for drawing samples from appropriate distributions and then correcting those draws to better target the posterior distribution. MCMC methods sample from the posterior distribution, which avoids computing complex partial first and second derivatives of the log likelihood function when a high number of dimensions are present. The Metropolis–Hastings algorithm samples from the joint posterior distribution, $P(q, b|X)$ —that is, sampling from the joint posterior probability distribution for vectors of ability parameters, q , item parameters, b , given data, X . The properties of these MCMC estimates were investigated in a series of simulation studies that evaluated the degree of parameter recovery (Yao, 2003, 2005). Because the Metropolis–Hastings method is used, moments for priors and proposals must be assigned and the variance/covariance between

ability factors. A two-step estimation process is also available, in which parameter estimates from a previous run are used as start values.

In the program, there is less reliance on the visual monitoring of the stationary process for the MCMC in **BMIRT** unlike generalized implementations such as **WINBUGS**. A simple **R** program is supplied that can be used to monitor the MCMC process. The User's Manual provides some brief guidance on MCMC convergence issues. The author suggests that parameters for the prior and proposals should be adjusted to obtain acceptance rates between 20% and 40% during the MCMC sampling.

To get some idea concerning the computational time required and capability to estimate a model, a very large and sparse operational data matrix was used that is computationally challenging. This data set consisted of slightly more than 41,000 students and 383 items that were a mixture of multiple-choice and constructed-response using the M-2PL and M-2PPC models. A two-factor exploratory model using 1,000 cycles required 42 min while a confirmatory four-dimensional model required approximately 45 min to estimate on a 32-bit operating system. Data processing can easily be handled using batch files. However, the real accomplishment is that large, somewhat ill-conditioned operational data sets can be estimated with high-dimensional structures using **BMIRT**.

Program Installation, Documentation, and Operating Environment

The program is free software and can be downloaded from www.BMIRT.com. Program documentation is provided, and substantive examples are given for a variety of problems. Technical explanation is given for some of the models and MCMC procedures. In many instances, the illustrative examples provided with the program can be modified to suit the data set under consideration.

BMIRT is implemented under the **JAVA** runtime environment. The program control language can be used to create, easily modify, and run multiple sequential analyses in batch files. There is no limit to the number of items or examinees that can be used with the program. The sample size required for calibration is consistent with those made for unidimensional applications. Four types of files are required to implement the **BMIRT** program: *BMIRT.bat*, *sample.ctf*, *sample.rwo*, and *sample.out*. The batch file (*.bat) calls the program control file (*.ctf). The control file, in turn, contains program setup information including the numbers of examinees, population groups, items, dimensions, iterations and burn-in cycles, as well as priors and specifications for the proposal distributions, item types, and patterns of item-factor loadings. The priors and proposals used in the example files can be used as a starting point and modified as necessary. In general, the default ones used in the example files work reasonably well. The variances and covariances for the latent ability parameters, q , must also be specified for all population groups.

The program has the capability to terminate the MCMC cycles after a fixed number of iterations and automatically use those final estimates as starting values for the next analysis. A set number of iterations are specified for the burn-in. Different MCMC chains can be run using different user defined start values and compared. The "rwo" file contains the examinee by item data, and the "out" files contain various types of MIRT statistics using a defined file name. Table 1 summarizes the available **BMIRT** output that is in American Standard Code for Information Interchange (ASCII) fixed format. The **PLINK** program implemented for **R** (Weeks, 2010) can read the **BMIRT** item parameter output to produce secondary item analysis such as item characteristic curves.

Table 1. Synopsis of BMIRT Output Files.

.param.txt	MCMC sampling for item parameters
.theta.txt	MCMC sampling for all examinee ability after burn-in
.likelihood.txt	MCMC iteration results (–log likelihood functions)
.par	Final item parameters β
.ss	Final examinees ability θ
.dm	Population mean μ
.dv	Population variance–covariance matrix σ
.AIC	Model fit statistics
.lerror	MCMC error of item parameters
.Aerror	MCMC error of ability parameters
.ScoreDistribution	Group score distribution (final item and ability estimates)
.ScoreDis	Score distribution for each group, MCMC samples of 50
.domainscore	Domain scores
.posterior.mean	Mean of the population posterior distribution
.posterior.var	Variance–covariance of the population posterior distribution
.item.sd	Item parameter estimates and variance matrix for MCMC errors
.LR.txt	Lord chi-square test and the degrees of freedom
.BF.txt	Model fit statistics

Note. BMIRT = Bayesian Multivariate Item Response Theory; MCMC = Markov chain Monte Carlo.

Conclusion

BMIRT is a powerful multipurpose program for multivariate IRT that can be used to address a wide variety of problems. Research should compare **BMIRT** with a variety of other MIRT-capable software, including **IRTPRO** (Cai, Thissen, & du Toit, 2011) and **EQSIRT** (Wu & Bentler, 2011). Simulations that compare the technical properties of estimates from these programs should be undertaken. With the recognition that the program is offered free with some good working examples provided, the number and types of models afforded has outstripped the available supporting documentation. Of course, the need for additional program documentation is not unique when considering specialized programs of this type.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Chicago, IL: SSI International.
- Grima, A. M., & Yao, L. (2011, April). *Classification consistency and accuracy: Unidimensional versus multidimensional IRT procedures*. Paper presented at the annual meetings of the National Council on Measurement in Education, New Orleans, LA.
- Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement*, 16, 279-293.

- Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item response theory. *Applied Measurement in Education*, 5, 193-211.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Weeks, J. P. (2010). PLINK: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, 35. Retrieved from <http://www.jstatsoft.org/v35/i12/>
- Wu, E. J., & Bentler, P. M. (2011). EQSIRT: A user-friendly IRT program [Computer software]. Encino, CA: Multivariate Software.
- Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [Computer software]. Monterey, CA: Defense Manpower Data Center.
- Yao, L. (2005). *An investigation of scaling options in estimating parameters for multidimensional item response theory model*. Unpublished manuscript.
- Yao, L. (2014). Subscore reporting through multidimensional IRT modeling. In Y. Cheng & H.-H. Chang (Eds.), *Advancing methodologies to support both summative and formative assessments*. Charlotte, NC: Information Age Publishing.
- Yao, L., & Li, F. (2010, May). *A DIF detection procedure in multidimensional framework and its applications*. Paper presented at the annual meetings of the National Council on Measurement in Education, Denver, CO.