# Fast High-Dimension Large-Scale IRT Factor Analysis for Polytomous Data

## Introduction

The estimation of item and person parameters in the framework of item response theory (IRT) is a common practice with regard to educational and psychological assessments. More recently, there has been great interest in estimation procedures for multidimensional IRT (MIRT). There is a wide range of applications for such models both in the field of education, but for other fields as well. While many assessments are developed according to a theoretical blueprint, it is often the case that the empirical item structure diverges from the expected structure. Thus, exploratory, as opposed to confirmatory, MIRT analysis is an important tool for critiquing assessments as well as for the theoretical development of the constructs targeted by those assessments.

The goal of this paper is to compare two different estimation procedures for IRT factor analysis of polytomous data. Two data sets are used. One is small with 734 cases and 24 items; one is a large 10-dimensional simulation with 100 items and 100,000 examinees. The first estimation procedure is Metropolis-Hastings Robbins-Monro algorithm for factor analysis (Cai, 2010) which is the core function of the MHRM estimation option in the software flexMIRT (Houts & Cai, 2016). The second is the SAEM algorithm adapted for factor analysis (Camilli & Geis, under review; Camilli & Fox, 2015). In this paper, it is shown that the SAEM method’s implementation results in a significant increase in efficiency, shaving more than 40% off the computation time of flexMIRT.

## Modeling Framework

Bock and Aitken (1981) in a seminal paper had proposed an expectation maximization [EM] algorithm based on the work of Dempster, Laird, and Rubin (1977), and extended this procedure to multidimensional item-response data. The computational inefficiencies for marginalizing across missing variables as the number of dimensions increases (the curse of dimensionality) presented a major obstacle for estimation, especially with high-dimensional data or a large number of variables.

A major trend in addressing computational challenges is the implementation of stochastic approaches to EM (e.g. Albert, 1992; Meng & Schilling, 1996; Cai, 2010) for marginalization. The goal of the approach within this paper is to demonstrate an implementation of an alternative EM algorithm for factor analysis based on stochastic approximation EM [SAEM] (Delyon, Lavielle, and Moulines, 1999). The concept of missing data as introduced by Tanner and Wong (1987) provided an important link between Monte Carlo methods and the EM algorithm. Béguin and Glas (2001) extended the Gibbs sampling approach of Albert (1992) to MIRT factor models, as well as stochastic versions of the Newton-Raphson algorithm (Gu & Kong, 1998; Cai, 2010) in which Metropolis-Hastings [MH] sampling is combined with the Robbins-Monro procedure [MH-RM] for establishing convergence (Robbins & Monro, 1951). Below, an improved algorithm is briefly presented for the estimation of coefficients in MIRT models relying on the foundational work of Meng and Schilling (1996) who proposed a Gibbs sampling approach to the factor analysis of dichotomous variables based on sufficient statistics.

The underlying model is identical for the two approaches to estimation (MHRM and SAEM), both for dichotomous and polytomous items. Assume a set of test items  for subjects , where the items are dichotomously scored: correct responses are scored  and incorrect responses . Assume there is a set of *Q* latent variables that account for an examinee’s observed item responses. For mathematical convenience, the cumulative normal function (also termed the normal ogive) is used rather than the logistic. These functions give virtually the same results in most applications; however, normal ogive models provide a convenient tool for the stochastic method shown below.

The basis of the two parameter normal ogive function (2PNO) for a correct response on item *j* presented to examinee *i*. Using the cumulative normal distribution function , the multidimensional probability for dichotomous responses (without guessing) is given by

 (1)

whereis a *Q*×*1* vector of latent factor scores or abilities for examinee *i*, is slope parameter, and is a guessing parameter. Discrimination for item *j* is now generalized to a *1*×*Q* vector of slopes or *factor loadings*. The factor loadings signify that an examinee’s response may be affected by *Q* different skills or abilities.

In MIRT, the underlying abilities  are considered to be latent variables. Person parameters are assumed to be independent of the item characteristics. A latent propensity can be used to construct a computational MIRT model, the latent space is conceptualized as a multivariate regression of the item propensities (or utilities)on:

 (2)

In terms of notation, for a *J*-item test, each row of **A** and **b** represent the slopes (discriminations) and difficulties for item *j*, is a  vector of factor scores, and  is a random measurement error.

For polytomous responses, assume there are *K* ordinal categories of an observed item response. The latent regression model is given by

 (3)

whereindexes decentered item thresholds and measurement errors. Note the latent propensity formulations were proposed by Albert (1992) and Albert and Chib (1993).

3. Estimation Procedures

The first estimation procedure is that is that implemented in the IRT software flexMIRT (Houts & Cai, 2016) in which a complete description of the estimation process is given. Note we say process because the MHRM algorithm is implemented at stage 3 of the estimation procedure. This provides a benchmark as it can be implemented within the same operating system on the same hardware so as to provide some insight into the utility and computational efficiency of the authors’ application of the SAEM algorithm.

In the technical appendix, the algorithm for polytomous data is described; it builds on previous work (Bock & Aitken, 1981; Takane & de Leeuw, 1986; Albert, 1992; Meng & Schilling, 1996; Fox, 2003; Cai, 2010) and the SAEM estimation procedure is employed (Delyon et al., 1999; Kuhn & Lavielle, 2005). This implementation is shown for polytomous items of which the procedure for dichotomous items is a special case. The convenience of the SAEM algorithm is based on the fact that the can be used to compute sufficient statistics for item parameters in the latent space (but not for the observed data!). On successive cycles, the latent sufficient statistics can be updated, and after burn in, can be subjected to the Robbin-Monro implementation of gain constant. This innovative strategy complements the general procedure of Meng and Schilling (1996) for carrying out the M-step of the EM algorithm with latent sufficient statistics.

The main cost of the SAEM procedure described in the appendix is random number generation, but the following features of the software speeded up calculations. First, in generating random numbers, we capitalized on R’s unique handling of matrices which allows vectorizing over items (columns of length *m*) and people (rows of length *n*) simultaneously, where n is sample size and m is number of items. The vectorized functions can then be executed in parallel with the R snow function parSapply in Windows or the R parallel function mcapply in Linux. Second, the data and item parameters were exported to common memory (clusterExport) for parallel analysis. Third, we found the R function mvnfast to significantly outperform the MASS function for the generation of MVN random variables. We note that the SAEM procedure in the paper for polytomous items is programmed entirely in R and requires less than 75 lines of code.

4. Data

In this section, the MHRM and SAEM methods are applied to polytomous data for a tryout form of 24 polytomous items (five categories) assessing the social quality of life for the Pediatric Quality of Life scale. These data for 753 children were previous analyzed by Cai (2010). Cai obtained solutions for 1-5 dimensions, and compared two methods (EM with adaptive quadrature, MHRM) in terms of factor structure and computation time. For a 5-factor solution, Cai (2010) found the EM algorithm required over 50 times more CPU time than MHRM (87 minutes v. 95 seconds.

5. Results

We compared the updated MHRM (Houts & Cai, 2016) and SAEM methods for a 3-factor solution. We found the SAEM algorithm (using parSapply) required 13 seconds of CPU time to converge at 1e-3 while the procedure (including MHRM) employed in flexMIRT required 12 seconds. We then optimized the SAEM estimation code by additional vectorization smaller size, reducing memory load, and compressing dependent operations into fewer lines of code, but using the basic algorithm. The optimized code required 9.37 seconds of CPU time v. about 12 seconds in flexMIRT.

6. Significance of work

We nonetheless believe it is too early to rank the estimation procedures in terms of CPU time at this stage of research. Variations in processing times is dependent upon resources available for professional programming, computing platforms, and operating systems (not to mention the bells and whistles of estimation parameters). Rather, the argument for the SAEM algorithm is that computing times are roughly comparable to other stochastic methods and the algorithm is extremely easy to employ in R by anyone who has written code for a Gibbs sampler. This opens up possibilities for wider implementation of exploratory IRT factor analysis. It is also the case that random number generation in the SAEM is a classic “embarrassingly parallel” problem. This sets the stage for outsourcing random number generation to a GPU board, and R should have this capacity in the near future. We expect estimation time can be substantially reduced allowing the practical application of IRT factor analysis to situations with 100s of variables and 100,000s of cases.

SAEM factor analysis is a topic worthy of additional investigation. The argument is not that SAEM provides a superior alternative to the procedure of Houts & Cai (2016), but rather that SAEM may open up new possibilities with modern multi-processor computing resources. A feature shared both by flexMIRT and the SAEM estimation procedure is that processing time goes up minimally for additional numbers of factors, while the corresponding time for adaptive quadrature goes up geometrically.

An additional capacity that is well underway is the factor analysis of large-scale assessment data bases such as TIMSS (Trends in Mathematics and Science Study) and NAEP (National Assessment of Educational Progress). One goal of factor analysis in this context is to construct empirical subscores that can be used to complement traditional subscores that are based on test design. While a total score is useful as a “horserace” indicator, it is too coarse of a tool for profiling jurisdictions, and the correlations among traditional subscores, when corrected for unreliability, typically approach *r* = 1 at the country or state level. As summarized by Wainer and Feinberg (2015, p. 18) “While it is too early to say that there are no subscores that are ever reported that are worth having, it seems sensible that unless tests are massively redesigned such subscores are likely rare.”

For an example of an application, consider the NAEP data jurisdiction-level (usually state or city) data. From this information, factor analysis could be used to obtain empirical subscores that have lower inter-correlation than tradition subscores. Such scores can be obtained for profiling jurisdictions—which have different academic strengths and weakness from an achievement perspective. A more useful goal would be to develop diagnostic information at the school level. However, factor analysis of large-scale data sets would provide policy analysts with better information for studying the impacts of educational policies on a large-scale.

## Technical Appendix

Let and  be a set of missing latent variables and *y* is the response data:



where represents a fixed value of the item parameters. The complete data likelihood is

,

where the multivariate variance term is  which constitutes an identification restriction. Note the last term in represents the normal prior  which has the effect of fixing the latent scale for ability. As shown by Béguin and Glass (2001), the values in can be sampled from a truncated normal distribution and  can be sampled as



Maximizing  with respect to **A** and **b** then results in





The SAEM algorithm’s computational efficiency results from the use of sufficient statistics of latent samples within each Gibbs cycle to estimate:



The conditional statistics depend on two sufficient statistics of the missing data:



This suggests a simple strategy based wholly on :





where **U** is the identity matrix containing the variances of the latent measurement errors *ε*. While the solution in and requires further identification restrictions, e.g. the lower triangular restriction of Anderson and Rubin (1956), the eigenanalysis of  is adequate for the identification of **A**.

Assume *κ* ordinal categories of an observed item response. The latent regression model is given by:



where  indexes item thresholds.

1. *Draw missing values x for estimating item thresholds*.

Ordinal item option propensities  are drawn as



Random values of  are independently generated for each individual for each item from the truncated normal distributions in . Item difficulties  and decentered thresholds  are obtained from sufficient statistics based on the *x* variables



2. *Draw missing values z for estimating ability*.

For dichotomous items and so  because there is just one item propensity. For  draw *z* from the truncated normal distribution

` 

where  and 

3. *Update sufficient statistics.*



****

where is the current value of the Robbins-Monro gain coefficient. The second equation specifically refers to the update to the augmented data covariance .

4. *Obtain*  **b***, and item thresholds from  and *

5. *Sample .*

6. *Repeat Steps 1-5 until convergence*.

In Step 2, note that convergence is defined relative to the sufficient statistics rather than values of fixed item parameters. Due to rotational indeterminacy individual fixed parameters such as factor loadings should not be monitored (except in simulations) during iterations.

To monitor convergence of parameters, a window size (say *W* = 3) is selected along with a tolerance constant. Iterations are terminated when the maximum covariance change for the trace of is less than  for *W* iterations. See Houts and Cai (2017) for an example of this convergence strategy with respect to individual parameters during MH-RM iterations. Robbins-Monro iterations require a defined step size; let , where  ensures convergence (Delyon et al., 1999).More details on this procedure can be found in Camilli and Fox (2015) and Camilli and Geis (under review).

References

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

Albert, J. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669-679.

Béguin, A. A. & Glas, C.A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66, 541-562.

Bock, R. D. & Aitkin, M. (1981), Marginal maximum likelihood estimation of item parameters: An application of a EM algorithm. *Psychometrika* , 46, 179-197.

Camilli, G. & Fox, J.-P. (2015). An aggregate IRT procedure for exploratory factor analysis. *Journal of Educational and Behavioral Statistics*, 40, 377-401.

Camilli, G. & Geis, E. (under review). Stochastic Approximation EM for Exploratory Item Factor Analysis.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33-57.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27, 94-128.

Houts, C. R., & Cai, L. (2016). *flexMIRT user’s manual version 3.0*: *Flexible multilevel item factor analysis and test scoring*. Seattle, WA: Vector Psychometric Group.

Meng, X.L. & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, 91, 1254-1267.

Wainer, H. & Feinberg, R. (2015). For want of a nail: Why unnecessarily long tests may be impeding the progress of Western civilization*. Significance*, 12 (1), 16-21.