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

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).

2. Modeling framework

The model is identical for the two approaches to estimation (MHRM and SAEM), both for dichotomous and polytomous items. Assume a set of test items $j = 1...J$ for subjects $i = 1...N$, where the items are dichotomously scored: correct responses are scored $Y_{ij} = 1$ and incorrect responses $Y_{ij} = 0$. 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 three parameter normal ogive function (3PNO) for a correct response on item j presented to examinee i. Using the cumulative normal distribution function Φ, the multidimensional probability for dichotomous responses (without guessing) is given by

$$P_{ij}\left(Y_{ij} = 1\right) = \Phi\left(Y_{ij} = 1 \mid \eta_{ij}\right)$$
$$\eta_{ij} = \mathbf{A}_j \boldsymbol{\theta}_i - b_j,$$

(1)

where $\boldsymbol{\theta}_i$ is a $Q \times 1$ vector of latent factor scores or abilities for examinee i, b_j is slope parameter, and c_j is a guessing parameter. Discrimination for item j is now generalized to $\mathbf{A}_j$, a $1 \times Q$ 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 $\boldsymbol{\theta}$ 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) z on $\boldsymbol{\theta}$:

$$
\begin{aligned}
z_{ij} &= \mathbf{A}_j \boldsymbol{\theta}_i - b_j + \varepsilon_{ij} \\
&= \eta_{ij} + \varepsilon_{ij}.
\end{aligned}
\tag{2}
$$

In terms of notation, for a *J*-item test, each row of **A** and **b** represent the slopes (discriminations) and difficulties for item *j*, $\boldsymbol{\theta}_i \sim N(\mathbf{0}, \mathbf{I})$ is a $Q \times 1$ vector of factor scores, and $\varepsilon_{ij} \sim N(0,1)$ 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

$$
\begin{aligned}
x_{ijk} &= A_j \theta_i - b_j - \tau_{jk} + \varepsilon_{ijk} \\
&= \eta_{ij} - \tau_{jk} + \varepsilon_{ijk}.
\end{aligned}
\tag{3}
$$

where $k = 1, 2, ..., K-1$ indexes 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. In the technical appendix, the SAEM algorithm for polytomous data is described. The convenience of the SAEM algorithm is based on the fact that the x_{ijk} 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.

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

The Gibbs sampling approach of Béguin, A. A. and Glas (2001) is expressed in terms of the complete data log likelihood:

$$
\begin{aligned}
l\left(\xi \mid z, \theta\right) &\propto -\frac{1}{2}\sum_i \sum_j \left\{ \left[z_{ij} - \left(\mathbf{A}_j \boldsymbol{\theta}_i - b_j \right) \right]^2 + \sum_q \theta_{iq}^2 \right\} \\
&= -\frac{1}{2}\sum_i \left\{ \left[\mathbf{z}_i - \left(\mathbf{A}\boldsymbol{\theta}_i - \mathbf{b} \right) \right]^T \left[\mathbf{z}_i - \left(\mathbf{A}\boldsymbol{\theta}_i - \mathbf{b} \right) \right] + \boldsymbol{\theta}_i^T \boldsymbol{\theta}_i \right\}.
\end{aligned}
\tag{4}
$$

Note the last term in (4) represents the normal prior $\boldsymbol{\theta} \sim N\left(0, \mathbf{I} \right)$. Building on previous work (Bock & Aitken, 1982; Takane & de Leew, 1986; Albert, 1992; Meng & Schilling, 1996), the estimation procedure in this paper then applies the SAEM method of Delyon et al. (1999). The *t*th iteration of this algorithm is:

1. Draw augmented data

   Sample $\boldsymbol{\theta}^{(t)}$ from $p\left(\boldsymbol{\theta} \mid \mathbf{z}^{(t-1)}, \xi^{(t-1)} \right)$

   Sample $\mathbf{z}^{(t)}$ from $p\left(\mathbf{z} \mid \boldsymbol{\theta}^{(t)}, \xi^{(t-1)} \right)$

2. Update sufficient statistic estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of complete data posterior

$$
\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \gamma_t \left\{ \overline{\mathbf{z}}^{(t)} - \boldsymbol{\mu}^{(t-1)} \right\}
$$

$$
\boldsymbol{\Sigma}^{(t)} = \boldsymbol{\Sigma}^{(t-1)} + \gamma_t \mathbf{S}_z^{(t)}, \text{ where } \mathbf{S}_z^{(t)} = \text{cov}\left(\mathbf{z}^{(t)} \right)
$$

3. Estimate item parameters

$$
\mathbf{b}^{(t)} = \arg \max_b \; l\left(\xi \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} \right)
$$

$$
\mathbf{A}^{(t)} = \arg \max_A \; l\left(\xi \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)} \right)
$$

For polytomous responses, random values of x_{ijk} are generated for each individual for each item from the truncated normal distribution. Similar to the setup for dichotomous items (see Camilli & Fox, 2015), the complete data for the log likelihood for polytomous items is

$$
\begin{aligned}
l\left(\xi \mid x, \theta\right) &= -\frac{1}{2}\sum_i \sum_j \sum_k \left[\left(x_{ijk} + \tau_{jk}\right) - \eta_{ij}\right]^2 \\
&= -\frac{1}{2}\sum_{ijk}\left[\left(x_{ijk} + \tau_{jk}\right)^2 - 2\left(x_{ijk} + \tau_{jk}\right)\eta_{ij} + \eta_{ij}^2\right].
\end{aligned}
\tag{6}
$$

Random values of x_{ijk} are generated for each individual for each item from the truncated normal distribution. This leads to the complete data log likelihood

$$
\begin{aligned}
l\left(\xi \mid x, \theta\right) &= -\frac{1}{2}\sum_i \sum_j \sum_k \left[\left(x_{ijk} + \tau_{jk}\right) - \eta_{ij}\right]^2 \\
&= -\frac{1}{2}\sum_{ijk}\left[\left(x_{ijk} + \tau_{jk}\right)^2 - 2\left(x_{ijk} + \tau_{jk}\right)\eta_{ij} + \eta_{ij}^2\right].
\end{aligned}
\tag{7}
$$

The third and fourth terms of the likelihood do not involve $\boldsymbol{\theta}$. Taking the derivative of the expected value of the likelihood w.r.t. b_j and τ_{jk}, respectively, and setting these expressions to zero results in

$$
\begin{aligned}
b_j &= -\sum_i \bar{x}_{ij\bullet}/n \\
\tau_{jk} &= -\sum_i \left(x_{ijk} - \bar{x}_{ij\bullet}\right)/n.
\end{aligned}
\tag{8}
$$

This indicates the sufficient statistics are $\sum_i \bar{x}_{ij\bullet}/n$ for item difficulties b_j, and $\sum_i \left(x_{ijk} - \bar{x}_{ij\bullet}\right)/n$ for decentered thresholds. To estimate $\mathbf{A}$, the sufficient statistic is $\operatorname{cov}\left(\bar{\mathbf{x}}_{\cdot}\right)$, and factor loadings can be obtained with an eigenanalysis optimizer.

Let $\mathbf{V}^{-1} = diag\left(K - 1\right)$ for a set of polytomous items all having K options. The ML solution for factor score is given by

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_i &\sim \left(\mathbf{I} + \mathbf{A}^T \mathbf{V}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^T \mathbf{V}^{-1}\left(\bar{\mathbf{x}}_i + \mathbf{b}\right) \\
&= \left[\left(\mathbf{I} + \left(K - 1\right)\mathbf{A}^T \mathbf{A}\right)^{-1}\right]\left[\left(K - 1\right)\mathbf{A}^T\left(\bar{\mathbf{x}}_i^T + \mathbf{b}\right)\right].
\end{aligned}
\tag{9}
$$

A Gibbs step within each iteration is used to draw augmented data as

$$
\boldsymbol{\theta}_i \sim MVN\left[\hat{\boldsymbol{\theta}}_i, \left(I + \left(K - 1\right)\mathbf{A}^T\mathbf{A}\right)^{-1}\right].
\tag{10}
$$

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.