PSYCHOMETRIKA—VOL. 75, NO. 1, 33–57
MARCH 2010
DOI: 10.1007/S11336-009-9136-X

HIGH-DIMENSIONAL EXPLORATORY ITEM FACTOR ANALYSIS BY A METROPOLIS–HASTINGS ROBBINS–MONRO ALGORITHM

LI CAI

UNIVERSITY OF CALIFORNIA, LOS ANGELES

A Metropolis–Hastings Robbins–Monro (MH-RM) algorithm for high-dimensional maximum marginal likelihood exploratory item factor analysis is proposed. The sequence of estimates from the MH-RM algorithm converges with probability one to the maximum likelihood solution. Details on the computer implementation of this algorithm are provided. The accuracy of the proposed algorithm is demonstrated with simulations. As an illustration, the proposed algorithm is applied to explore the factor structure underlying a new quality of life scale for children. It is shown that when the dimensionality is high, MH-RM has advantages over existing methods such as numerical quadrature based EM algorithm. Extensions of the algorithm to other modeling frameworks are discussed.

Key words: stochastic approximation, SA, item response theory, IRT, Markov chain Monte Carlo, MCMC, numerical integration, categorical factor analysis, latent variable modeling, structural equation modeling.

1. Introduction

Full-information Item Factor Analysis (IFA; Bock, Gibbons, & Muraki, 1988) has long been a useful tool for exploring the latent structure underlying educational and psychological tests. It is also being increasing utilized in mental health and quality of life research due to a recent surge of interest among researchers in the application of item response theory to develop standardized measurement instruments for patient reported outcomes. A notable example is the National Institutes of Health Patient Reported Outcomes Measurement Information System (PROMIS; Reeve, Hays, Bjorner, Cook, Crane, & Teresi, 2007). IFA proves to be crucial in these new domains of application, yet despite recent advances in methods for fitting high-dimensional item response theory models, maximum marginal likelihood estimation in IFA remains a difficult numerical problem. The biggest obstacle stems from the need to evaluate intractable high-dimensional integrals in the likelihood function for the item parameters. Depending on how the integrals are approximated, existing algorithms for maximum likelihood estimation in IFA can be grouped roughly into the following four classes.

The first class involves adaptive Gaussian quadrature. By replacing fixed-point quadrature rules with adaptive rules (Liu & Pierce, 1994; Naylor & Smith, 1982), approximations to the high-dimensional integrals impose significantly less computational burden. Adapting the quadrature nodes also stabilizes likelihood computations, because when the number of items is large

I thank the editor, the AE, and the reviewers for helpful suggestions. I am indebted to Drs. Chuanshu Ji, Robert MacCallum, and Zhengyuan Zhu for helpful discussions. I would also like to thank Drs. Mike Edwards and David Thissen for supplying the data sets used in the numerical demonstrations. The author gratefully acknowledges financial support from Educational Testing Service (the Gulliksen Psychometric Research Fellowship program), National Science Foundation (SES-0717941), National Center for Research on Evaluation, Standards and Student Testing (CRESST) through award R305A050004 from the US Department of Education's Institute of Education Sciences (IES), and a predoctoral advanced quantitative methods training grant awarded to the UCLA Departments of Education and Psychology from IES. The views expressed in this paper are of the author's alone and do not reflect the views or policies of the funding agencies.

Requests for reprints should be sent to Li Cai, GSE & IS, UCLA, Los Angeles, CA, USA 90095-1521. E-mail: lcai@ucla.edu

the likelihood becomes so concentrated that standard Gaussian quadrature formulae do not accurately capture its mass. With care in implementation, pointwise convergence of the estimates to a local maximum of the likelihood function can be obtained (e.g., Rabe-Hesketh, Skrondal, & Pickles, 2005; Schilling & Bock, 2005). Because of over two decades of success with Bock and Aitkin's (1981) EM algorithm, adaptive quadrature based EM algorithm is often considered a gold standard against which other algorithms are compared. It is also possible to use adaptive quadrature in a Newton–Raphson algorithm such as in GLLAMM (Rabe-Hesketh, Skrondal, & Pickles, 2004a). Despite its popularity, adaptive quadrature still limits the number of factors that an IFA software can handle simply because the number of quadrature points must grow exponentially as the dimensionality of the latent traits increases. This phenomenon is often referred to as the "curse of dimensionality" in the literature. In addition, if the EM algorithm is used in conjunction with quadrature, variability information of parameter estimates is not an automatic by-product. Additional computation for parameter standard errors is required (see, e.g., Cai, 2008b) upon EM's convergence. As a result, TESTFACT does not print standard errors in its output.

The second class is characterized by the use of Laplace approximation (Tierney & Kadane, 1986). Applications of this method can be found in Kass and Steffey (1989), Thomas (1993), and Huber, Ronchetti, and Victoria-Feser (2004). In the context of IFA, the Laplace method is essentially adaptive Gauss–Hermite integration with 1 quadrature point. It is computationally fast (see, e.g., Raudenbush, Yang, & Yosef, 2000, in a slightly different application), but a notable feature of this method is that the error of approximation decreases only as the number of items increases. When few items are administered to each examinee, such as in an adaptive test design, or when there are relatively few items loading on a factor, such as in the presence of testlets (Wainer & Kiely, 1987), the degree of imprecision in approximation can become substantial and may lead to biased parameter estimates (Joe, 2008). Raudenbush et al. (2000) argue for the use of higher-order Laplace approximation, but the complexity of software implementation grows dramatically as the order of approximation increases. In addition, the truncation point in the asymptotic series expansion (6th degree in their paper) of the integrand function is essentially arbitrary. Thus, the utility of the Laplace method in high-dimensional full-information IFA remains an open question.

The third class of methods is intimately related to Wei and Tanner's (1990) MCEM algorithm, wherein Monte Carlo integration replaces numerical quadrature in the E-step (e.g., Meng & Schilling, 1996; Song & Lee, 2005). To achieve pointwise convergence, simulation size (the number of random draws for Monte Carlo integration) must increase as the estimates move closer to the maximum so that Monte Carlo error in the E-step does not overwhelm changes in the M-step. To automate the amount of increase in simulation size, adaptive algorithms have been devised (e.g., Booth & Hobert, 1999), but the number of random draws in the final iterations of these adaptive algorithms can become prohibitively high (in the order of tens of thousands as observed by Jank, 2004), dramatically slowing down MCEM's convergence. The MCEM algorithm is also inefficient in the use of simulated data because at each E-step, a new set of random draws are generated, and all previous draws are discarded.

The fourth class is purely stochastic. A defining characteristic of this class of algorithms is the use of fully Bayesian sampling-based estimation methods such as Markov chain Monte Carlo (MCMC; Tierney, 1994). Within the Bayesian estimation framework, maximum likelihood can be approximated by choosing an appropriate non-informative prior distribution. Since properties of the posterior distribution of the item parameters are of primary interest, one constructs an ergodic Markov chain whose unique invariant measure is the posterior, and then after a certain "burn-in" period, samples from the chain may be regarded as random draws from the posterior, from which any functional of the posterior distribution can be estimated. While the basic principle is easy to state, the implementations vary to a wide extent (Albert, 1992; Béguin & Glas, 2001; Dunson, 2000; Patz & Junker, 1999a, 1999b; Segall, 1998; Shi & Lee,

1998), and the relative algorithmic efficiency of the existing implementations have not been entirely settled (see, e.g., Edwards, 2005). Prior specification (particularly of the noninformative kind) is another inherent difficulty (see, e.g., Natarajan & Kass, 2000).

From the preceding discussion, it seems clear that a flexible and efficient algorithm that converges pointwise to the maximum likelihood estimate (MLE) is much desired for high-dimensional IFA. Indeed, in the research proposed here, a Metropolis–Hastings Robbins–Monro (MH-RM) algorithm is suggested to address most of the afore-mentioned difficulties. The MH-RM algorithm is well suited to general computer programming for large-scale analysis involving many items, many factors, and many respondents. It is efficient in the use of Monte Carlo because the simulation size is fixed and usually small throughout the iterations. In addition, it also produces an estimate of the parameter information matrix as a by-product that can be used subsequently for standard error estimation and goodness-of-fit testing (e.g., Cai, Maydeu-Olivares, Coffman & Thissen, 2006).

In brief, the MH-RM algorithm is a data augmented Robbins–Monro type (RM; Robbins & Monro, 1951) stochastic approximation (SA) algorithm driven by the random imputations produced by a Metropolis–Hastings sampler (MH; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The MH-RM algorithm is motivated by Titterington's (1984) recursive algorithm for incomplete data estimation, and is a close relative of Gu and Kong's (1998) SA algorithm. It can also be conceived of as a natural extension of the Stochastic Approximation EM algorithm (SAEM; Celeux & Diebolt, 1991; Celeux, Chauveau, & Diebolt, 1995; Delyon, Lavielle, & Moulines, 1999). Probability one convergence of the sequence of estimates to a local maximum of the likelihood surface is established along essentially the same line as Gu and Kong's (1998) Theorem 1.

SA algorithms have been well studied in the fields of systems engineering, adaptive control, and signal processing (see, e.g., Benveniste, Métivier, & Priouret, 1990; Borkar, 2008; Kushner & Yin, 1997) since the pioneering work of Robbins and Monro (1951). Until recently, statistical applications of SA algorithms have remained predominantly in the area of generalized and nonlinear mixed-effects modeling (Gu & Kong, 1998; Gu & Zhu, 2001; Gu, Sun, & Huang, 2004; Gueorguieva & Agresti, 2001; Kuhn & Lavielle, 2005; Makowski & Lavielle, 2006; Zhu & Lee, 2002). While the IFA model can be thought of as a nonlinear mixed model, it has features requiring specialized software implementation for practical testing situations.

The remainder of this paper is organized as follows. First, an IFA model for graded responses is introduced in Section 2. The MH-RM algorithm is derived in Section 3. Section 4 addresses details for efficiently implementing MH-RM for IFA and compares MH-RM with the Bock and Aitkin (1981) EM algorithm by means of a small simulation study. Section 5 contains results from two empirical studies in which MH-RM is compared with quadrature based EM algorithm. It is shown that MH-RM has distinct advantages in terms of speed, stability, and flexibility. Extensions to the basic MH-RM algorithm is discussed in Section 6, and the paper concludes with directions for future research in Section 7.

2. A Model for Item Factor Analysis

2.1. A Multidimensional Graded Model

This section (re)introduces notation for a logistic IFA model for graded responses. The derivations are straightforward extensions of Samejima's (1996) graded response model and bears some similarity to the multidimensional model of te Marvelde, Glas, and van Damme (2006). Let there be $i = 1, \ldots, N$ independent respondents, $j = 1, \ldots, n$ items. For item j, let there be C_j response categories. Let y_{ij} denote the response from respondent i to item j. Suppose there are p factors and $\boldsymbol{\beta}_j$ is the $p \times 1$ vector of item slopes for item j, and $\mathbf{x}_i$ is the

$p \times 1$ vector of factor scores for respondent i. As is customarily assumed, the factor scores follow a multivariate normal distribution with a null mean vector and identity covariance matrix. Let $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{j(C_j-1)})'$ be a $(C_j - 1) \times 1$ vector of category intercepts for item j. Let $\boldsymbol{\theta}_j = (\boldsymbol{\alpha}_j', \boldsymbol{\beta}_j')'$ be a vector containing all parameters for item j. Conditional on the item parameters and $\mathbf{x}_i$, define the following set of boundary response probabilities:

$$\begin{aligned} P(y_{ij} \geq 0|\boldsymbol{\theta}_j, \mathbf{x}_i) &= 1, \\ P(y_{ij} \geq 1|\boldsymbol{\theta}_j, \mathbf{x}_i) &= \frac{1}{1 + \exp(-\alpha_{j1} - \boldsymbol{\beta}_j'\mathbf{x}_i)}, \\ &\ldots \\ P(y_{ij} \geq C_j - 1|\boldsymbol{\theta}_j, \mathbf{x}_i) &= \frac{1}{1 + \exp(-\alpha_{j(C_j-1)} - \boldsymbol{\beta}_j'\mathbf{x}_i)}, \\ P(y_{ij} \geq C_j|\boldsymbol{\theta}_j, \mathbf{x}_i) &= 0. \end{aligned} \tag{1}$$

It follows that the conditional probability for the response $y_{ij} = k$ is given by

$$\pi_{ijk} = P(y_{ij} = k|\boldsymbol{\theta}_j, \mathbf{x}_i) = P(y_{ij} \geq k|\boldsymbol{\theta}_j, \mathbf{x}_i) - P(y_{ij} \geq k + 1|\boldsymbol{\theta}_j, \mathbf{x}_i), \tag{2}$$

for $k \in \{0, 1, \ldots, C_j - 1\}$. Note that not all parameters are identified (estimable) in this model. Reflection and rotation of the factor pattern are both possible. Identification can be achieved by fixing $p(p-1)/2$ slopes to zero. Rotation to simple structure is still necessary for the interpretation of the factor pattern.

2.2. *Observed and Complete Data Likelihood*

First, it is useful to define an indicator function

$$\chi_k(y) = \begin{cases} 1, & \text{if } y = k, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

for $k \in \{0, 1, \ldots, C_j - 1\}$. It follows from Equation (2) that the conditional distribution of y_{ij} is that of a multinomial with C_j cells, trial size 1, and cell probabilities π_{ijk}:

$$f(y_{ij}|\boldsymbol{\theta}_j, \mathbf{x}_i) = \prod_{k=0}^{C_j-1} \pi_{ijk}^{\chi_k(y_{ij})}. \tag{4}$$

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})'$ be the ith person's response pattern. By the conditional independence assumption (Lord & Novick, 1968), the conditional density of $\mathbf{y}_i$ is

$$f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i) = \prod_{j=1}^{n} f(y_{ij}|\boldsymbol{\theta}_j, \mathbf{x}_i), \tag{5}$$

where $\boldsymbol{\theta}$ is a $d \times 1$ parameter vector containing the estimable item parameters for all n items. For a person randomly sampled from a population with standard multivariate normally distributed latent traits, the marginal density of $\mathbf{y}_i$ is

$$f(\mathbf{y}_i|\boldsymbol{\theta}) = \int \prod_{j=1}^{n} f(y_{ij}|\boldsymbol{\theta}_j, \mathbf{x}) \boldsymbol{\Phi}(d\mathbf{x}), \tag{6}$$

where $\mathbf{\Phi}(\cdot)$ is the standard multivariate normal distribution function, and the integral in Equation (6) is a p-fold Lebesgue–Stieltjes integral over $\mathbb{R}^p$. Let $\mathbf{Y}$ be an $N \times n$ matrix of individual response patterns, whose ith row is $\mathbf{y}_i'$. The observed data likelihood is

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^{N}\left[\int \prod_{j=1}^{n} f(y_{ij}|\boldsymbol{\theta}_j, \mathbf{x})\mathbf{\Phi}(d\mathbf{x})\right]. \tag{7}$$

The factor scores can be thought of as missing data. Let $\mathbf{X}$ be an $N \times p$ matrix of factor scores whose ith row is $\mathbf{x}_i'$. The observed data $\mathbf{Y}$ can be augmented by missing data $\mathbf{X}$ to permit the representation of complete data as $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. The complete data likelihood for the IFA model has the following factored form

$$L(\boldsymbol{\theta}|\mathbf{Z}) = \prod_{i=1}^{N}\left[\boldsymbol{\phi}(\mathbf{x}_i)\prod_{j=1}^{n} f(y_{ij}|\boldsymbol{\theta}_j, \mathbf{x}_i)\right] = \left[\prod_{i=1}^{N}\boldsymbol{\phi}(\mathbf{x}_i)\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{n} f(y_{ij}|\boldsymbol{\theta}_j, \mathbf{x}_i)\right], \tag{8}$$

where $\boldsymbol{\phi}(\cdot)$ is the standard multivariate normal density.

2.3. MLE, Sparseness, and Goodness-of-Fit

Direct maximization of $L(\boldsymbol{\theta}|\mathbf{Y})$ in Equation (7) leads to the maximum marginal likelihood estimator $\hat{\boldsymbol{\theta}}$. As Bock and Aitkin (1981) showed, $L(\boldsymbol{\theta}|\mathbf{Y})$ is a multinomial likelihood function based on an underlying contingency table of the full cross-classifications of the item responses with $T = \prod_{j=1}^{n} C_j$ cells. Therefore, $\hat{\boldsymbol{\theta}}$ is referred to as the full-information estimator in the literature, in contrast with the limited-information estimators that identify the item parameters from lower-order marginal tables. A comparison of full-versus-limited information estimators for item factor analysis is beyond the scope of this paper. Interested readers are referred to the recent reviews by Bolt (2005) and Wirth and Edwards (2007). In brief, the full-information estimator is more flexible, can readily handle missing responses, and provides a basis for the development of Bayesian estimators (see, e.g., Mislevy, 1986). A large body of applied item response theory research related to educational and psychological testing relies on this estimator as implemented in computer programs such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) and MULTILOG (Thissen, 2003).

A feature of the contingency table considered here is that if the number of items n is large, this table becomes sparse—a well-known issue that Bartholomew and Knott (1999) discuss in detail. The MLE $\hat{\boldsymbol{\theta}}$ itself is root-N consistent and asymptotically normally distributed with minimum variance (Bishop, Fienberg, & Holland, 1975) for a given table size T. However, a major problem arises when one attempts to use full-information goodness-of-fit statistics such as the likelihood ratio G^2 or Pearson's X^2 statistic to test the absolute fit of the item factor model. The sparseness invalidates the use of the asymptotic chi-square approximation as the reference distribution for these statistics. Recent advances in limited-information goodness-of-fit testing have partly addressed this recurring difficulty (e.g. Bartholomew & Leung, 2002; Cai et al., 2006; Maydeu-Olivares & Joe, 2005). As to the likelihood ratio testing of two nested models, under conditions stated by Haberman (1977), sparseness does not invalidate the chi-square approximation for the likelihood ratio G^2 difference statistic if the larger (less restrictive one) of the two models is correct (see also Table 1 in Maydeu-Olivares & Cai, 2006). As to standard errors, results in Cai (2008b) show that when the table is sparse, the inverse of the information matrix continue to serve as a useful characterization of the asymptotic covariance matrix of the parameter estimates.

2.4. Factor Loadings in Normal Metric

The logistic item response model is preferred in maximum likelihood estimation due to simplifications in calculations of the log-likelihood derivatives (Baker & Kim, 2004). On the other hand, for historical reasons, exploratory factor analysis results are usually presented as a matrix of rotated factor loadings in the standardized normal metric. Thus, in keeping with the psychometric tradition, and to facilitate the reporting of comparative studies in Section 5 involving computer software with different parameterizations, the item parameters are converted into thresholds and loadings in normal metric. Formulae for such conversions are standard results and can be found in many places (e.g., Wirth & Edwards, 2007).

Central to the conversion is a scaling constant D that puts the logistic parameters on a normal ogive metric. Traditionally D is taken to be 1.702 (Camilli, 1994) based on the minimax principle, but recently Savalei (2006) derived a new scaling constant $D = 1.749$ from Kullback and Leibler's (1951) information criterion. The old constant $D = 1.702$ is used in the sequel to remain consistent with standard practice. Let $\boldsymbol{\alpha}_j^* = (1/D)\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j^* = (1/D)\boldsymbol{\beta}_j$. Then the thresholds $\boldsymbol{\tau}_j$ and factor loadings $\boldsymbol{\lambda}_j$ in normal metric can be computed as

$$\boldsymbol{\tau}_j = \frac{-\boldsymbol{\alpha}_j^*}{\sqrt{1 + (\boldsymbol{\beta}_j^*)'\boldsymbol{\beta}_j^*}}, \qquad \boldsymbol{\lambda}_j = \frac{\boldsymbol{\beta}_j^*}{\sqrt{1 + (\boldsymbol{\beta}_j^*)'\boldsymbol{\beta}_j^*}}. \tag{9}$$

3. A Metropolis-Hastings Robbins-Monro Algorithm

3.1. The EM Algorithm and Fisher's Identity

Using the notation of Section 2.2, where $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$, the complete data likelihood is $L(\boldsymbol{\theta}|\mathbf{Z})$ for a d-dimensional parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Suppose $\mathbf{X} \in \mathcal{E}$, where $\mathcal{E}$ is some sample space. The task is to compute the MLE $\hat{\boldsymbol{\theta}}$ based on the observed data likelihood $L(\boldsymbol{\theta}|\mathbf{Y})$.

Let $l(\boldsymbol{\theta}|\mathbf{Y}) = \log L(\boldsymbol{\theta}|\mathbf{Y})$ and $l(\boldsymbol{\theta}|\mathbf{Z}) = \log L(\boldsymbol{\theta}|\mathbf{Z})$. Instead of maximizing $l(\boldsymbol{\theta}|\mathbf{Y})$ directly, Dempster, Laird, and Rubin (1977) transformed the observed data estimation problem into a sequence of complete data estimation problems by iteratively maximizing the conditional expectation of $l(\boldsymbol{\theta}|\mathbf{Z})$ over $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})$, where $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})$ denotes the conditional distribution of missing data given observed data. Let the current estimate be $\boldsymbol{\theta}^*$. One iteration of the EM algorithm consists of: (a) the E(xpectation) step, in which the expected complete-data log-likelihood is computed as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = \int_{\mathcal{E}} l(\boldsymbol{\theta}|\mathbf{Z})\Pi(d\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}^*), \tag{10}$$

and (b) the M(aximization)-step, in which $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$ is maximized to yield an updated estimate. Let

$$\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z}) = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Z}) \tag{11}$$

be the gradient of the complete data log-likelihood, where $\nabla_{\boldsymbol{\theta}}$ returns a $d \times 1$ vector of first order derivatives of $l(\boldsymbol{\theta}|\mathbf{Z})$ with respect to $\boldsymbol{\theta}$. By Fisher's Identity (Fisher, 1925), the conditional expectation of $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})$ over $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})$ is equal to the gradient of the observed data log-likelihood:

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y}) = \int_{\mathcal{E}} \mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})\Pi(d\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}). \tag{12}$$

The MH-RM algorithm is strongly motivated by Fisher's Identity. Equation (12) suggests, rather counter-intuitively that in a gradient-based scheme, one can optimize $l(\boldsymbol{\theta}|\mathbf{Y})$ without directly evaluating its gradient. Instead, the ascent directions are given by the conditional expectation of the complete data gradient $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})$. A solution that is a zero for the right-hand side of (12) also satisfies the likelihood equations and is an optimizer of $l(\boldsymbol{\theta}|\mathbf{Y})$. The central connection lies in taking the expectation of $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})$ with respect to the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$, which amounts to augmenting missing data from its posterior predictive distribution. Since $\boldsymbol{\theta}$ is unknown and $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$, the solution can only be obtained iteratively. The MH-RM algorithm is no more than a formalization of this idea.

3.2. MH-RM as a Data Augmented RM Algorithm

Robbins and Monro's (1951) algorithm is a root-finding algorithm for noise-corrupted regression functions. In the simplest case, let $g(\cdot)$ be a real-valued function of a real variable θ. If $g(\cdot)$ were known and continuously differentiable, one can use Newton's procedure

$$\theta_{k+1} = \theta_k + \left[-\nabla_\theta g(\theta_k)\right]^{-1} g(\theta_k)$$

to find its root. Alternatively, if differentiability cannot be assumed, one can use the following successive approximation:

$$\theta_{k+1} = \theta_k + \gamma g(\theta_k)$$

in a neighborhood of the root if γ is sufficiently small. Now suppose that $g(\theta)$ can only be measured imprecisely as $g(\theta) + \zeta$, where ζ is a zero mean random variable representing the noise process. This is the original situation Robbins and Monro (1951) were dealing with. The Robbins–Monro method iteratively updates the approximation to the root according to the following recursive scheme:

$$\theta_{k+1} = \theta_k + \gamma_k R_{k+1}, \tag{13}$$

where $R_{k+1} = g(\theta_k) + \zeta_{k+1}$ is an estimate of $g(\theta_k)$ and $\{\gamma_k; k \geq 1\}$ is a sequence of *gain constants* such that:

$$\gamma_k \in (0, 1], \quad \sum_{k=1}^{\infty} \gamma_k = \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} \gamma_k^2 < \infty. \tag{14}$$

Taken together, the three conditions ensure that the gain constants decrease *slowly* to zero. The intuitive appeal of this algorithm is that R_{k+1} does not have to be highly accurate. This can be understood from the following: if θ_k is still far away from the root, taking a large number of observations to compute a good estimate of $g(\theta_k)$ is inefficient because R_{k+1} is useful insofar as it provides the right direction for the next move. The decaying gain constants eventually eliminate the noise effect so that the sequence of estimates converges to the root.

The MH-RM algorithm is an extension of the basic algorithm in Equation (13) to multiparameter problems that involve stochastic augmentation of missing data. Let

$$\mathbf{H}(\boldsymbol{\theta}|\mathbf{Z}) = -\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{Z})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

be the $d \times d$ complete data information matrix, and let $\mathcal{K}(\cdot, A|\mathbf{Y}, \boldsymbol{\theta})$ be a Markov transition kernel such that for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and any measurable set $A \in \mathcal{E}$, it generates a uniformly ergodic chain having $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})$ as its invariant measure so that

$$\int_A \Pi(d\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}) = \int_{\mathcal{E}} \Pi(d\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}) \mathcal{K}(\mathbf{X}, A|\mathbf{Y}, \boldsymbol{\theta}). \tag{15}$$

Let initial values be $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\Gamma}_0)$, where $\boldsymbol{\Gamma}_0$ is a $d \times d$ symmetric positive definite matrix. Let $\boldsymbol{\theta}^{(k)}$ be the parameter estimate at the end of iteration k. The $(k+1)$th iteration of the MH-RM algorithm consists of

- *Stochastic Imputation*: Draw m_k sets of missing data $\{\mathbf{X}_j^{(k+1)}; j = 1, \ldots, m_k\}$ from $\mathcal{K}(\cdot, A|\mathbf{Y}, \boldsymbol{\theta}^{(k)})$ to form m_k sets of complete data $\{\mathbf{Z}_j^{(k+1)} = (\mathbf{Y}, \mathbf{X}_j^{(k+1)}); j = 1, \ldots, m_k\}$. In practice, it is often useful to exploit the relation $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}) \propto L(\mathbf{Z}|\boldsymbol{\theta})$ and construct an MH sampler to produce these imputations.
- *Stochastic Approximation*: Using the relation in Equation (12), compute an approximation of $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}^{(k)}|\mathbf{Y})$ by the sample average of complete data gradients

$$\tilde{\mathbf{s}}_{k+1} = \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{s}\big(\boldsymbol{\theta}^{(k)}\big|\mathbf{Z}_j^{(k+1)}\big), \tag{16}$$

  and a recursive approximation of the conditional expectation of the complete data information matrix

$$\boldsymbol{\Gamma}_{k+1} = \boldsymbol{\Gamma}_k + \gamma_k \left\{ \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{H}\big(\boldsymbol{\theta}^{(k)}\big|\mathbf{Z}_j^{(k+1)}\big) - \boldsymbol{\Gamma}_k \right\}. \tag{17}$$

- *Robbins–Monro Update*: Set the new parameter estimate to

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \gamma_k \big(\boldsymbol{\Gamma}_{k+1}^{-1} \tilde{\mathbf{s}}_{k+1}\big). \tag{18}$$

The iterations are terminated when the estimates converge. In practice, γ_k may be taken as $1/k$, in which case the choice of $\boldsymbol{\Gamma}_0$ becomes arbitrary. One can show that under certain regularity conditions the MH-RM algorithm converges to a local maximum of $l(\boldsymbol{\theta}|\mathbf{Y})$ with probability one (see Appendix A). Though the simulation size m_k is allowed to depend on the iteration number k, it is by no means required. The convergence result shows that the algorithm converges with a fixed and relatively small simulation size, i.e., $m_k \equiv m$, for all k.

The MH-RM for maximum likelihood estimation is not too different from the engineering application of the RM algorithm for the identification and control of a dynamical system with observational noise. Finding the MLE amounts to finding the root of $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$, but because of missing data, $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$ is difficult to evaluate directly. In contrast, the gradient of the complete data log-likelihood $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})$ is often much simpler. Making use of Fisher's identity in Equation (12), the conditional expectation of $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})$ is equal to $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$, so if one can augment missing data by sampling from a Markov chain having $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})$ as its target, $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$ can be approximated by taking a sample average, as in Equation (16).

As to the matrix $\boldsymbol{\Gamma}_k$, it approximates the conditional expectation of $\mathbf{H}(\boldsymbol{\theta}|\mathbf{Z})$ over $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})$. In multiparameter optimization, use of curvature information often speeds up convergence. The complete data information matrix is easy to compute, especially so in IFA (see Section 4.2), and the recursive filter in Equation (17) helps stabilize the Monte Carlo noise. The term $(\boldsymbol{\Gamma}_{k+1}^{-1} \tilde{\mathbf{s}}_{k+1})$ serves precisely the same role as R_{k+1} in Equation (13). Finally, in Equation (18), MH-RM proceeds by using the same recursive filter as Equation (13) to average out the effect of the simulation noise on parameter estimates, so that the sequence of estimates converges to the root of $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$ with probability one.

3.3. Relation of MH-RM to Some Existing Algorithms

It is easy to see that stochastic imputation in MH-RM replaces deterministic Gaussian quadrature in Bock and Aitkin's (1981) EM algorithm. By doing so MH-RM escapes from the

"curse of dimensionality." One can also understand MH-RM from the angle of Joint Maximum Likelihood (JML; see Baker & Kim, 2004)—a historically popular estimator in IRT. JML computations iterate between two stages that are similar to the first and last stages in MH-RM: (1) replacing the unobserved factor scores with modal estimates given current item parameters, and (2) maximizing the log-likelihoods of the items with factor scores treated as known. JML is not necessarily convergent because a single modal estimate fails to acknowledge the inherent uncertainty due to not observing the factor scores, whereas the variability of the stochastic imputations in MH-RM ensures that this uncertainty is properly accounted for.

Cai (2006) showed that when the complete data log-likelihood corresponds to that of the generalized linear model for exponential family outcomes, the MH-RM algorithm can be derived as an extension of the SAEM algorithm by the same linearization argument that leads to the iteratively reweighted least squares algorithm (McCullagh & Nelder, 1989) for maximum likelihood estimation in generalized linear models. This result implies that if the complete data model is ordinary multiple linear regression for Gaussian outcomes (e.g., conventional linear factor analysis), the SAEM algorithm and the MH-RM algorithm are numerically equivalent. In other cases when this finite-time numeric equivalence does not hold, Delyon et al. (1999) showed that the SAEM algorithm has the same asymptotic (in time) behavior as the stochastic gradient scheme. Equation (18) makes it clear that the MH-RM algorithm is a stochastic gradient algorithm, which implies that MH-RM and SAEM share the same asymptotic dynamics.

The MH-RM algorithm has much in common with Gu and Kong's (1998) stochastic approximation Newton–Raphson algorithm. However, the two algorithms differ in an important way. Gu and Kong's (1998) algorithm uses an estimate of the information matrix of the observed data log-likelihood whereas MH-RM uses the conditional expectation of the complete data information matrix. By the missing information principle (Orchard & Woodbury, 1972), the step size of the MH-RM algorithm is smaller than Gu and Kong's (1998) algorithm. As it will become clear in Section 4.2, by making smaller step sizes, the MH-RM algorithm becomes easier to implement, requires much less computation per iteration, and is more stable than Gu and Kong's (1998) algorithm whenever the complete data likelihood is of a factored form. This will subsequently be important because the IFA model has a factored complete data likelihood.

If one sets γ_k to be identically equal to unity throughout the iterations, the MH-RM algorithm becomes a Monte Carlo Newton–Raphson algorithm (MCNR; McCulloch & Searle, 2001). Unlike MCEM, there is no explicit maximization step in the MH-RM algorithm, so the two are not transparently related. However, if $\gamma_k \equiv 1$, the Robbins–Monro update step can be thought of as a single iteration of maximization, in the same spirit as Lange's (1995) algorithm with a single iteration of Newton–Raphson in the M-step, which turns out to be locally equivalent to the EM algorithm. Thus, the MH-RM algorithm with constant step size may be taken as a stochastic counterpart of Lange's (1995) gradient algorithm. MH-RM is also closely related to Titterington's (1984) algorithm for incomplete data estimation.

In addition to γ_k being unity, if the number of iterations is also equal to one, i.e., $m_k \equiv 1$ for all k, the MH-RM algorithm becomes a close relative of Diebolt and Ip's (1996) stochastic EM (SEM) algorithm. The sequence of estimates produced by the SEM algorithm forms a time-homogeneous Markov chain. The mean of its invariant distribution is close to the MLE, and the variance reflects loss of information due to missing data. In psychometric models similar to IFA, the SEM algorithm is found to converge quickly to a close vicinity of the MLE (see, e.g., Fox, 2003). Thus, the version of MH-RM similar to the SEM algorithm leads to a simple and effective method for computing start values for the subsequent MH-RM iterations with decreasing gain constants. The implementation details will be elaborated in Section 4.3.

3.4. Approximating the Information Matrix

Following Louis (1982), the information matrix of the observed data log-likelihood is

$$-\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{Y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \left[\int_{\mathcal{E}} \mathbf{H}(\boldsymbol{\theta}|\mathbf{Z})\Pi(d\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}) - \int_{\mathcal{E}} \mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})\big[\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})\big]'\Pi(d\mathbf{X}|\mathbf{Y},\boldsymbol{\theta})\right] + \int_{\mathcal{E}} \mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})\Pi(d\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}) \int_{\mathcal{E}} \big[\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})\big]'\Pi(d\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}).$$

This is a direct consequence of Orchard and Woodbury's (1972) missing information principle. Note that the part in the square brackets can be recursively approximated as

$$\boldsymbol{\Psi}_k = \boldsymbol{\Psi}_{k-1} + \gamma_k\left\{\frac{1}{m_k}\sum_{j=1}^{m_k}\big[\mathbf{H}\big(\boldsymbol{\theta}^{(k)}\big|\mathbf{Z}_j^{(k)}\big) - \mathbf{s}\big(\boldsymbol{\theta}^{(k)}\big|\mathbf{Z}_j^{(k)}\big)\big[\mathbf{s}\big(\boldsymbol{\theta}^{(k)}\big|\mathbf{Z}_j^{(k)}\big)\big]'\big] - \boldsymbol{\Psi}_{k-1}\right\},$$

and Fisher's identity in Equation (12) suggests the following procedure to recursively approximate the score vector:

$$\boldsymbol{\psi}_k = \boldsymbol{\psi}_{k-1} + \gamma_k\left\{\frac{1}{m_k}\sum_{j=1}^{m_k}\mathbf{s}\big(\boldsymbol{\theta}^{(k)}\big|\mathbf{Z}_j^{(k)}\big) - \boldsymbol{\psi}_{k-1}\right\}.$$

Putting the pieces together, the observed data information matrix can be approximated as

$$\mathcal{I}_k = \boldsymbol{\Psi}_k - \boldsymbol{\psi}_k\boldsymbol{\psi}_k'. \tag{19}$$

As k tends to infinity and the MH-RM iterations converge,

$$\mathcal{I}_k \to -\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{Y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}.$$

The inverse of $\mathcal{I}_k$ is the large-sample covariance matrix of the parameter estimates.

4. Implementing the MH-RM Algorithm for IFA

4.1. The MCMC Imputation Procedure

The MCMC procedure for imputing the factor scores can be derived in a similar way as in Patz and Junker (1999a) from a Metropolis-within-Gibbs calculation (Chib & Greenberg, 1995). Let $\xi(\mathbf{x}_i|\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\ldots,\mathbf{x}_N,\mathbf{Y},\boldsymbol{\theta})$ be the full conditional density for $\mathbf{x}_i$, and let $\mathbf{x}_i^l$ be the value of $\mathbf{x}_i$ in the lth iteration of a Gibbs sampler consisting of the following steps:

$$\begin{aligned}
&\text{Draw } \mathbf{x}_1^l \sim \xi\big(\mathbf{x}_1\big|\mathbf{x}_2^{(l-1)},\ldots,\mathbf{x}_N^{(l-1)},\mathbf{Y},\boldsymbol{\theta}\big),\\
&\text{Draw } \mathbf{x}_2^l \sim \xi\big(\mathbf{x}_2\big|\mathbf{x}_1^l,\mathbf{x}_3^{(l-1)},\ldots,\mathbf{x}_N^{(l-1)},\mathbf{Y},\boldsymbol{\theta}\big),\\
&\qquad\cdots\\
&\text{Draw } \mathbf{x}_i^l \sim \xi\big(\mathbf{x}_i\big|\mathbf{x}_1^l,\ldots,\mathbf{x}_{i-1}^l,\mathbf{x}_{i+1}^{(l-1)},\ldots,\mathbf{x}_N^{(l-1)},\mathbf{Y},\boldsymbol{\theta}\big),\\
&\qquad\cdots\\
&\text{Draw } \mathbf{x}_N^l \sim \xi\big(\mathbf{x}_N\big|\mathbf{x}_1^l,\ldots,\mathbf{x}_{N-1}^l,\mathbf{Y},\boldsymbol{\theta}\big).
\end{aligned} \tag{20}$$

Let the transition kernel defined by this Gibbs sampler be $\mathcal{K}(\mathbf{X}, A|\mathbf{Y}, \boldsymbol{\theta})$. Standard results (e.g., Gelfand & Smith, 1990) ensure that it satisfies the condition in Equation (15). Hence if $\mathbf{X}_l = \{\mathbf{x}_i^l; i = 1, \ldots, N\}$, the sequence $\{\mathbf{X}_l; l \geq 0\}$ converges in distribution to $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})$.

The full conditionals are difficult to directly sample from, but they are specified up to a proportionality constant, i.e.,

$$\begin{aligned}&\xi(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_N, \mathbf{Y}, \boldsymbol{\theta})\\ &\quad\propto L(\boldsymbol{\theta}|\mathbf{Z}) = f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)\left[\prod_{h\neq i}^{N}\boldsymbol{\phi}(\mathbf{x}_h)\prod_{j=1}^{n} f(y_{hj}|\boldsymbol{\theta}_j, \mathbf{x}_h)\right].\end{aligned}$$

This suggests coupling the Gibbs sampler with the MH algorithm. To draw each $\mathbf{x}_i$, the following MH transition kernel is used:

$$\begin{aligned}&\mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\mathbf{Y}, \boldsymbol{\theta})\\ &= q(\mathbf{x}_i, \mathbf{x}_i^*)\min\left\{\frac{f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i^*)\boldsymbol{\phi}(\mathbf{x}_i^*)[\prod_{h\neq i}^{N}\boldsymbol{\phi}(\mathbf{x}_h)\prod_{j=1}^{n} f(y_{hj}|\boldsymbol{\theta}_j, \mathbf{x}_h)]q(\mathbf{x}_i^*, \mathbf{x}_i)}{f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)[\prod_{h\neq i}^{N}\boldsymbol{\phi}(\mathbf{x}_h)\prod_{j=1}^{n} f(y_{hj}|\boldsymbol{\theta}_j, \mathbf{x}_h)]q(\mathbf{x}_i, \mathbf{x}_i^*)}, 1\right\}d\mathbf{x}_i^*\\ &= q(\mathbf{x}_i, \mathbf{x}_i^*)\min\left\{\frac{f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i^*)\boldsymbol{\phi}(\mathbf{x}_i^*)q(\mathbf{x}_i^*, \mathbf{x}_i)}{f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)q(\mathbf{x}_i, \mathbf{x}_i^*)}, 1\right\}d\mathbf{x}_i^*\\ &= \mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\mathbf{y}_i, \boldsymbol{\theta})\end{aligned} \tag{21}$$

for $\mathbf{x}_i^* \neq \mathbf{x}_i$ and $\mathcal{K}(\mathbf{x}_i, \{\mathbf{x}_i\}|\mathbf{y}_i, \boldsymbol{\theta}) = 1 - \int_{\mathbf{x}_i^*\neq\mathbf{x}_i}\mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\mathbf{y}_i, \boldsymbol{\theta})$, where $q(\mathbf{x}_i, \mathbf{x}_i^*)$ is any aperiodic recurrent transition density. Piecing the Gibbs part and the MH part together, the transition kernel for generating the stochastic imputations can be written as

$$\mathcal{K}(\mathbf{X}, d\mathbf{X}^*|\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{N}\mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\mathbf{y}_i, \boldsymbol{\theta}). \tag{22}$$

In the sequel, a simple random walk chain $\mathbf{x}_i^* = \mathbf{x}_i + \mathbf{e}_i$ is used to generate the proposal draws, where the increment density is that of a scaled standard multivariate normal distribution in p dimensions, i.e., $\mathbf{e}_i \sim \mathcal{N}_p(\mathbf{0}, c^2\mathbf{I}_p)$. The scalar parameter c adjusts the dispersion of the increments, so one can change its value to tune the acceptance ratio of the MH chain. Simple calculation shows that $q(\mathbf{x}_i, \mathbf{x}_i^*) = \det(2\pi c^2\mathbf{I}_p)\exp\{-(\mathbf{x}_i^* - \mathbf{x}_i)'(\mathbf{x}_i^* - \mathbf{x}_i)/(2c^2)\}$ for this increment density. Because $q(\mathbf{x}_i, \mathbf{x}_i^*) = q(\mathbf{x}_i^*, \mathbf{x}_i)$, Equation (21) can be further reduced to

$$\mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\mathbf{y}_i, \boldsymbol{\theta}) = q(\mathbf{x}_i, \mathbf{x}_i^*)\min\left\{\frac{f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i^*)\boldsymbol{\phi}(\mathbf{x}_i^*)}{f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)}, 1\right\}d\mathbf{x}_i^*. \tag{23}$$

The kernel in Equation (22) represents a remarkably simple sampling plan because all the conditioning kernels $\mathcal{K}(\mathbf{x}_i, d\mathbf{x}_i^*|\mathbf{y}_i, \boldsymbol{\theta})$ on the right-hand side can be evaluated independently of each other. This means that the N updates in Equation (20) can be finished simultaneously, if a matrix-oriented programming language such as GAUSS (Aptech Systems, Inc., 2003) is used. In brief, one first generates an $N \times p$ matrix $\mathbf{E}$, whose ith row is $\mathbf{e}_i'$, from a matrix normal distribution (Mardia, Kent, & Bibby, 1979) with independent rows each distributed as $\mathcal{N}_p(\mathbf{0}, c^2\mathbf{I}_p)$, and compute the proposals as $\mathbf{X}^* = \mathbf{X} + \mathbf{E}$. Then for all rows, one evaluates the acceptance probabilities in Equation (23). Because the acceptance probability for each new $\mathbf{x}_i$ only depends on the item response part of the IFA model $f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{x}_i)$ for that particular i, the acceptance probabilities can be computed as a "dot" division of two vectors.

Since $\mathbf{Y}$ is fixed, let $\mathcal{K}_k(\cdot, A) = \mathcal{K}(\cdot, A|\mathbf{Y}, \boldsymbol{\theta}^{(k)})$ denote the transition kernel in the $(k+1)$th iteration of MH-RM. From initial state $\mathbf{X}_0^{(k)}$, a sequence $\{\mathbf{X}_l^{(k)}; l \geq 0\}$ is generated by iterating $\mathcal{K}_k(\mathbf{X}, A)$, i.e.,

$$\Pr\big(\mathbf{X}_l^{(k)} \in A \big| \mathbf{X}_0^{(k)}\big) = \mathcal{K}_k^l\big(\mathbf{X}_0^{(k)}, A\big),$$

where $\mathcal{K}_k^l(\mathbf{X}_0^{(k)}, A)$ denotes the lth iterate of the kernel. The sequence of random imputations $\{\mathbf{X}_j^{(k)}; j = 1, \ldots, m_k\}$ can be chosen from $\{\mathbf{X}_l^{(k)}; l \geq 0\}$ as a subsequence, using standard "burn-in" and/or "thinning" methods. The initial state can be chosen as the last element of $\{\mathbf{X}_j^{(k-1)}; j = 1, \ldots, m_{k-1}\}$, i.e., $\mathbf{X}_0^{(k)} = \mathbf{X}_{m_{k-1}}^{(k-1)}$. One should tweak the scalar dispersion parameter c so that the rejection rates of the MH chain is within a reasonable range of the optimal rates as discussed by Roberts and Rosenthal (2001). While the current method may not have the optimal proposal distribution, it is simple to implement and does not involve a large amount of computation per iteration. As will be shown in Section 5, the performance of the MH chain is quite admirable. Note that standard subsampling methods have little impact on the asymptotic behavior of the MH-RM algorithm because the convergence proof does not require uncorrelated imputations. If the starting values are sufficiently close to the MLE, one may take $m_k \equiv 1$ for all k and set the number of "burn-in" iterates to as small as 5.

4.2. Complete Data Log-Likelihood and Derivatives

Using Equation (8), dropping parts that are constant, the complete data log-likelihood for the IFA model can be written as

$$l(\boldsymbol{\theta}|\mathbf{Z}) \propto \sum_{j=1}^{n}\sum_{i=1}^{N} \log f(y_{ij}|\boldsymbol{\theta}_j, \mathbf{x}_i) = \sum_{j=1}^{n}\left[\sum_{i=1}^{N}\sum_{k=0}^{C_j-1} \chi_k(y_{ij}) \log \pi_{ijk}\right]. \tag{24}$$

The part in the square brackets can be recognized as the log-likelihood for ordinal logistic regression (McCullagh, 1980). For the jth item, the imputed values in $\mathbf{X}$ serve as a matrix of predictors and the vector of observed responses in the jth column of $\mathbf{Y}$ is the outcome variable.

Next, the derivatives of the complete data model are needed. Note that in Equation (24) the ordinal regression models are independent of each other, so it suffices to consider a generic item j. The independence implies that the information matrix $\mathbf{H}(\boldsymbol{\theta}|\mathbf{Z})$ is block diagonal, with n blocks each corresponding to an item. For each item, the contributions to the log-likelihood and its derivatives are then summed over i.

The computational efficiency of the MH-RM algorithm over Gu and Kong's (1998) algorithm becomes evident, especially in the many-factor many-item case. The complete data log-likelihood in Equation (24) is a sum of n independent terms, so the derivative computation and parameter updating can be performed for each item separately, or even in parallel if the computing environment supports multiple processors. In addition, if C is the maximum number of categories across all items, the matrix inversion needed in MH-RM is at most of dimension $p + C - 1$ (cf. n times $(p + C - 1)$ in Gu and Kong's algorithm). The necessary first and second order derivatives for the IFA model are given in Appendix B.

4.3. Starting Values and Convergence Check

A two-stage procedure is used to find starting values for the MH-RM algorithm. First, an unweighted least squares factor extraction using the sample polychoric correlation matrix gives initial values to start M SEM-type iterations, wherein both the gain constants and the m_k's are

set to 1 for $k = 1, \ldots, M$. At the end of the Mth iteration, the sequence of parameter estimates obtained from SEM-type iterations are averaged and used as starting values for the subsequent MH-RM iterations with decreasing gain constants.

As Borkar (2008) noted, stability is a great virtue of stochastic approximation based algorithms. It makes small changes per cycle so that the algorithm has a graceful behavior. It is, of course, not infallible. If starting values are poor, or if the data contain little information about some of the parameters, the MH-RM algorithm will fail. However, the use of the two-stage starting value procedure appears to mitigate commonly encountered difficulties associated with EM-type algorithms using numerical quadrature (see, e.g., Section 5.2).

Convergence of the MH-RM can be monitored by computing a window of successive differences in parameter estimates. The iterations are terminated if and only if all differences in the window are less than some prescribed threshold, which is set to equal 1.0×10^{-4} in the computer applications reported in this paper. The window size is set to 3 to prevent premature stop due to random variation.

4.4. A Small Simulation Study

As a verification of the theoretical development, a small Monte Carlo simulation study was conducted to compare the MH-RM algorithm with the celebrated Bock and Aitkin (1981) EM algorithm in terms of parameter recovery and sampling variability. The data generating model is the IFA model presented in Section 2.1 with $p = 2$ dimensions and $n = 10$ three-category items. Table 1 presents the true item intercepts and slopes (in the logistic metric). These generating parameters were chosen to resemble values encountered in real data analysis. The number of Monte Carlo replications is 100 and the sample size N is equal to 1000. The relatively small number of replications is more than sufficient for the stated purposes.

In each replication, the 2-dimensional model (same as the data generating model) is fitted to the simulated data set, using both Bock and Aitkin (1981) EM and MH-RM. Both algorithms were implemented in the C++ programming language as part of the numeric engine in the prototype IRTPRO program (Cai, du Toit, & Thissen, 2009). For MH-RM, the simulation size m_k is set to equal 1 for all cycles and the gain constant $\gamma_k = 1/k$. For Bock and Aitkin (1981) EM, there are 20 quadrature points per dimension, equally spaced between -4 and 4. The EM iterations are deemed converged if the maximum absolute inter-iteration change in parameter estimates drops below 1.0×10^{-5}. For MH-RM, the convergence criteria are as given in Section 4.3. The starting values for both algorithms are: .25 for the first intercept for all items, $-.25$ for the second intercept for all items, and 1.0 for all slopes but the second slope of the first item, which is constrained to zero as part of the identification restrictions.

Both algorithms lead to converged solutions in all replications. Table 2 compares the estimated raw bias for estimating each parameter using Bock and Aitkin (1981) EM and MH-RM. Estimated raw bias is equal to the average of the difference between the estimates and the true parameter values over the Monte Carlo replications. As one can easily tell, both algorithms perform well in recovering the parameters and are nearly indistinguishable when measured by bias alone.

TABLE 1.
Simulation study: generating parameter values.

	Item									
	1	2	3	4	5	6	7	8	9	10
Intercept 1	.67	1.09	−.18	−.76	.50	−.41	−.07	1.15	.13	−1.10
Intercept 2	−.72	−.14	−1.22	−1.42	−.36	−1.26	−.96	−.09	−.70	−1.56
Slope 1	2.20	2.00	2.60	1.60	1.70	1.80	1.80	1.90	1.60	1.70
Slope 2	.00	.00	.00	.00	.00	1.20	1.10	1.20	2.10	1.50

TABLE 2.
Simulation study: raw bias (Monte Carlo standard deviations).

Item	Intercept 1		Intercept 2		Slope 1		Slope 2	
	BAEM	MHRM	BAEM	MHRM	BAEM	MHRM	BAEM	MHRM
1	.00 (.13)	−.01 (.13)	−.02 (.12)	−.02 (.12)	.03 (.20)	.02(.20)	N/A	N/A
2	.01 (.13)	.00 (.13)	.00 (.12)	.00 (.12)	.02 (.17)	.02(.17)	−.01 (.15)	−.01 (.14)
3	−.01 (.13)	−.01 (.13)	.00 (.14)	.00 (.14)	−.01 (.21)	−.01(.21)	.04 (.18)	.03 (.17)
4	−.01 (.09)	−.01 (.09)	−.01 (.11)	−.01 (.11)	.00 (.13)	.00(.13)	.00 (.12)	−.01 (.11)
5	−.01 (.10)	−.01 (.10)	.00 (.10)	−.01 (.10)	.01 (.14)	.01(.14)	.00 (.13)	.00 (.12)
6	.00 (.12)	−.01 (.12)	−.03 (.14)	−.03 (.14)	.01 (.17)	.01(.17)	.03 (.16)	.02 (.16)
7	−.01 (.11)	−.01 (.11)	−.01 (.10)	−.01 (.10)	.00 (.14)	.00(.14)	−.01 (.16)	−.01 (.15)
8	.00 (.13)	.00 (.13)	.00 (.12)	.00 (.12)	−.01 (.16)	−.01(.16)	.02 (.16)	.01 (.15)
9	.00 (.13)	.00 (.13)	.00 (.13)	−.01 (.13)	−.01 (.19)	.00(.18)	.02 (.30)	.03 (.31)
10	−.02 (.13)	−.02 (.13)	−.02 (.13)	−.02 (.13)	.02 (.18)	.03 (.18)	.02 (.21)	.02 (.21)

Note. BAEM = Bock–Aitkin EM algorithm; Bias is defined as the Monte Carlo average of estimates minus the corresponding true value; Monte Carlo Standard Deviations are in the parentheses. Slope 2 of item 1 is fixed to 0 as part of the identification constraints.

Also contained in Table 2 is the information about sampling variability. Specifically, the Monte Carlo standard deviations, defined as the observed standard deviations of the estimates over Monte Carlo replications, are again nearly indistinguishable between Bock and Aitkin (1981) EM and MH-RM. The total root mean square deviation from true values for all parameters is equal to .014 for both algorithms. The average of per replication run time is 30 seconds for Bock and Aitkin (1981) EM and 41 seconds for MH-RM. The simulation suggests that the maximum likelihood solutions produced by the Bock and Aitkin (1981) EM algorithm and the MH-RM algorithm have comparable quality, though for sufficiently low-dimensional problems, a fully optimized deterministic algorithm such as EM can be more efficient.

5. Numerical Illustrations

To examine the empirical performance of MH-RM, two sets of data were analyzed in a comparison of the proposed algorithm with well-established alternatives that include a Gibbs sampling based MCMC estimation algorithm (Section 5.1) and an adaptive Gauss–Hermite quadrature based EM algorithm (Section 5.2). To ensure fairness, only compiled native-code software programs written in a high-level language such as FORTRAN or C++ are used in the comparison. For MH-RM, the prototype IRTPRO (Cai et al., 2009) is used. The Gibbs sampling based MCMC method developed and implemented by Edwards (2005) is chosen because the C++ program (MultiNorm) was specifically designed for item factor analysis. The adaptive quadrature EM module in IRTPRO (Cai et al., 2009) is also used because of its exclusive focus on item factor analysis. While Mplus 5.0 (Muthén & Muthén, 2008) is less focused on item factor analysis, it serves as an independent benchmark in comparisons involving EM due to its wide availability. CEFA (Browne, Cudeck, Tateneni, & Mels, 2008) is used for factor rotation. All analyses were conducted on a laptop computer with a 2 GHz Intel Duo Core CPU and 2 GB of RAM. Parallel processing on multi-core or hyper-threaded CPUs is directly supported in IRTPRO and Mplus, but the capability is turned off in the comparisons.

5.1. MCMC for a Simulated Data Set

The data are the responses of 2000 simulees to a hypothetical scale consisting of 19 4-category graded items. There are four correlated factors underlying the item responses. The

TABLE 3.
Rotated factor loadings for simulated data with generating values as the target.

Item	Factor 1			Factor 2			Factor 3			Factor 4		
	Gen	MH	MC	Gen	MH	MC	Gen	MH	MC	Gen	MH	MC
1	.82	.87	.86	0	−.05	.00	0	.08	−.02	0	−.09	−.04
2	.84	.82	.85	0	.00	−.04	0	.02	.08	0	.03	−.05
3	.78	.76	.79	0	.06	−.01	0	−.02	.00	0	.00	.01
4	.69	.69	.67	0	.02	.04	0	−.07	−.04	0	.09	.07
5	.77	.75	.72	0	.03	.04	0	−.03	−.02	0	.03	.04
6	0	−.07	−.07	.81	.85	.85	0	−.05	−.04	0	.07	.05
7	0	.04	.05	.73	.71	.70	0	−.05	−.04	0	.04	.03
8	0	.08	.05	.65	.60	.63	0	.09	.04	0	−.14	−.07
9	0	.00	.00	.77	.78	.77	0	−.01	.01	0	−.03	−.04
10	0	−.02	−.02	.76	.73	.74	0	.06	.04	0	−.01	−.02
11	0	.02	.02	0	.02	.01	.86	.79	.79	0	.06	.05
12	0	−.02	−.01	0	.04	.03	.88	.83	.81	0	.05	.05
13	0	.01	.02	0	−.04	−.05	.80	.86	.86	0	−.05	−.05
14	0	−.01	−.01	0	.00	−.01	.74	.77	.78	0	−.03	−.03
15	0	−.03	−.03	0	.02	.02	.78	.77	.78	0	.03	.01
16	0	−.01	−.01	0	−.05	−.05	0	.06	.05	.78	.76	.76
17	0	−.01	.00	0	.04	.02	0	−.06	−.05	.78	.83	.81
18	0	−.03	−.02	0	.04	.01	0	.04	.05	.69	.62	.61
19	0	.10	.04	0	−.09	−.02	0	.03	−.02	.63	.57	.61

Note. Gen = Generating values serving as the target; MH = target rotated MH-RM estimates; MC = target rotated MCMC estimates.

TABLE 4.
Factor correlations after target rotation for simulated data.

	Factor Correlations (i, j)					
	(2, 1)	(3, 1)	(3, 2)	(4, 1)	(4, 2)	(4, 3)
Generating Value	.75	.70	.75	.60	.50	.80
MH-RM Estimates	.73	.67	.73	.62	.52	.79
MCMC Estimates	.74	.68	.75	.64	.56	.81

first three factors are each measured by 5 items and the last one by 4 items. The details for data simulation can be found in Edwards (2005) and Table 3 lists the generating factor pattern.

For MH-RM, the simulation size m_k is set to 1 and the gain constants $\gamma_k = 1/k$, with the same convergence criteria as given in Section 4.3. For the MCMC method, a total of 60,000 random draws were taken with diffuse priors on all item parameters. The first 10,000 draws are regarded as "burn-in" and the "thinning" interval is 50. As noted by Edwards (2005), these conservative choices ensure that the results are not dependent on the peculiarities of the starting values. The posterior means of the item parameters can be understood as approximate MLEs.

Given the availability of the generating parameters, a completely specified oblique target rotation (Browne, 2001) of the estimated factor loadings can be applied to both solutions, with the generating loadings serving as the target. A side-by-side comparison of the target rotated loadings are presented in Table 3, and a similar comparison of the factor correlations is available in Table 4. As can be seen, the MH-RM and MCMC estimates are very close to each other and both are close to the generating values. The Root Mean Square Deviation (RMSD) from target for MH-RM is 0.046 and the RMSD for MCMC is 0.039. It is worth noting that IRTPRO uses the logistic parameterization and MultiNorm uses the normal ogive parameterization. In addition, IRTPRO

TABLE 5.
Item wording for the social quality of life scale.

	Wording
Item 1	I could talk with my friends.
Item 2	I felt good about how I got along with classmates.
Item 3	I felt comfortable with other kids my age.
Item 4	I felt loved by my parents or guardians.
Item 5	I was good at making friends.
Item 6	Other kids wanted to be with me.
Item 7	I felt accepted by other kids my age.
Item 8	I did things with other kids my age.
Item 9	I was good at talking with adults.
Item 10	My teachers understood me.
Item 11	I wanted to spend time with my family.
Item 12	I had problems getting along with my parents or guardians.
Item 13	I got into a yelling fight with other kids.
Item 14	I had trouble getting along with my family.
Item 15	Other kids made fun of me.
Item 16	I felt bad about how I got along with my friends.
Item 17	I felt different from other kids my age.
Item 18	Other kids were mean to me.
Item 19	I felt nervous when I was with other kids my age.
Item 20	I did not want to be with other kids.
Item 21	I had trouble getting along with other kids my age.
Item 22	I got along better with adults than with other kids my age.
Item 23	I was afraid of other kids my age.
Item 24	I wished I had more friends.

explicitly optimizes a log-likelihood while MultiNorm does not. These numerical differences may have contributed to the .007 difference in RMSDs. There is, however, a large difference in computational time. MH-RM required 47 seconds of run time in 197 cycles and MCMC required 1 hour 20 minutes and 34 seconds of run time.

5.2. A Social Quality of Life Scale for Children

The data are the responses of 753 children (between the ages of 8 to 17) to 24 social quality of life items on an item tryout form for the Pediatric Quality of Life scales. Table 5 lists the text of item wording. The five-point response scale is "Never" (0), "Almost never" (1), "Sometimes" (2), "Often" (3), and "Almost Always" (4). The item responses have been recoded if necessary so that the highest numerical value of the response scale indicates positive social quality of life.

The adaptive quadrature based EM implementation in IRTPRO is used. The convergence criterion for EM considers a solution converged when the interiteration change in log-likelihood drops below .001. For MH-RM, the simulation size m_k is again set to 1 and the gain constants $\gamma_k = 1/k$, with the same convergence criteria as in the previous sections.

5.2.1. A Unidimensional Model. Initial calibration of the items using a unidimensional graded response model suggests that the model fits the data poorly and there is strong indication of local dependence. The unidimensional graded model is a special case of the graded IFA model with a single latent variable ($p = 1$). For numerical integration, 21 adaptive Gauss–Hermite quadrature points are used so that the log-likelihood can be approximated accurately.

Table 6 shows a side-by-side comparison of the two sets of item parameter estimates (in logistic metric) obtained from the two algorithms. The EM algorithm required 5 seconds of run

TABLE 6.
Unidimensional graded model parameter estimates.

Item	Intercept 1		Intercept 2		Intercept 3		Intercept 4		Slope (*SE*)	
	MH-RM	EM	MH-RM	EM	MH-RM	EM	MH-RM	EM	MH-RM	EM
1	3.32	3.32	2.60	2.61	.82	.82	−.09	−.08	1.31 (.10)	1.31 (.11)
2	3.07	3.07	2.19	2.19	.56	.56	−.59	−.59	.11 (.07)	.11 (.07)
3	3.42	3.42	2.70	2.71	.77	.77	−.40	−.39	.73 (.08)	.73 (.09)
4	4.53	4.53	3.72	3.72	1.88	1.89	.35	.36	1.62 (.13)	1.62 (.13)
5	3.28	3.28	2.58	2.59	1.03	1.04	−.39	−.38	1.32 (.10)	1.31 (.11)
6	3.42	3.42	2.46	2.46	.79	.79	−.41	−.40	.93 (.09)	.92 (.09)
7	4.14	4.13	3.14	3.14	1.10	1.10	−.60	−.59	1.96 (.13)	1.94 (.14)
8	3.45	3.45	2.66	2.66	1.13	1.13	.11	.12	.82 (.09)	.82 (.09)
9	3.58	3.59	2.94	2.95	1.54	1.56	.17	.18	1.67 (.12)	1.67 (.13)
10	4.04	4.05	3.77	3.78	2.74	2.74	1.62	1.63	.84 (.12)	.85 (.12)
11	4.65	4.65	3.91	3.91	2.65	2.66	1.33	1.34	1.22 (.13)	1.23 (.13)
12	4.22	4.23	3.30	3.30	1.70	1.70	.21	.21	1.71 (.12)	1.71 (.13)
13	3.88	3.88	2.71	2.72	1.07	1.08	−.05	−.04	1.70 (.12)	1.70 (.13)
14	2.52	2.52	1.60	1.60	−.28	−.27	−1.24	−1.24	.87(.08)	.87 (.09)
15	3.93	3.95	3.26	3.28	1.50	1.51	.19	.20	1.69 (.13)	1.70 (.13)
16	3.61	3.61	2.91	2.92	1.20	1.21	.12	.13	.91 (.09)	.91 (.09)
17	3.65	3.64	2.69	2.69	1.23	1.23	−.14	−.14	.71 (.09)	.70 (.09)
18	4.46	4.47	3.85	3.86	1.77	1.78	.26	.27	1.88 (.14)	1.88 (.14)
19	3.98	4.00	3.12	3.13	1.81	1.83	.47	.47	1.33 (.11)	1.34 (.12)
20	4.72	4.72	3.61	3.62	1.74	1.75	.33	.34	1.48 (.11)	1.48 (.12)
21	4.42	4.43	3.43	3.44	1.81	1.82	.55	.55	1.48 (.12)	1.49 (.13)
22	4.64	4.65	4.03	4.05	2.78	2.80	1.38	1.40	1.62 (.14)	1.63 (.15)
23	4.20	4.20	3.61	3.62	1.22	1.23	−.55	−.55	1.51 (.11)	1.51 (.12)
24	4.17	4.17	3.20	3.21	1.60	1.61	.04	.05	1.32 (.11)	1.32 (.11)

time in 47 cycles. The MH-RM algorithm required 10 seconds of run time in 128 cycles. As can be seen from Table 6, the two sets of estimates are nearly identical. The absolute difference between EM estimates and MH-RM estimates is no larger than .02. The standard errors of the slopes are also quite similar. The observed data log-likelihood is equal to −19590.4, according to EM. For MH-RM, the value is −19590.2. In this case there is no real difference in the quality of estimates. As expected, MH-RM is less efficient than EM for this unidimensional problem.

5.2.2. A Five-Dimensional Model. The combination of expert advice, the wording of the items in Table 5, as well as the existence of such problematic cases as Item #2 in the initial calibration suggests that there may well be additional dimensions underlying the 24-item scale. IFA is a useful tool for modeling these extra dimensions. Specifically, an IFA model with 5 latent variables in a pseudo-bifactor structure seems plausible.

The approach taken here is conventional. One first fits an exploratory IFA model with 5 factors to the data, and then rotates the loadings orthogonally to a partially-specified target (Browne, 2001). Plausibility of the hypothesized factor structure can be inferred from the RMSD of the rotated loadings from the target. Table 7 shows this target pattern. The entries in the table that are marked with an X indicate unspecified loadings whose magnitude is to be determined by rotation, whereas the 0s indicate loadings to be minimized by the target rotation. The first factor can be regarded as a primary social quality of life dimension. The other factors mainly account for extra local dependence.

Once again, both the MH-RM algorithm and the EM algorithm are used for fitting the IFA model. Due to an increase in the number of factors, the number of quadrature points per dimen-

TABLE 7.
Target rotated factor loadings.

Item	Factor 1			Factor 2			Factor 3			Factor 4			Factor 5		
	TP	MH	EM	TP	MH	EM	TP	MH	EM	TP	MH	EM	TP	MH	EM
1	X	.62	.61	0	.14	.14	0	−.12	−.10	0	.04	.01	0	.02	.02
2	X	.00	−.01	0	.19	.19	X	.40	.39	0	.18	.17	0	−.13	−.11
3	X	.28	.29	X	.38	.36	0	.09	.09	X	.32	.36	0	.06	.04
4	X	.57	.57	X	.47	.46	0	.01	.00	0	.16	.16	0	−.06	−.05
5	X	.61	.61	0	.08	.08	0	.02	.02	0	−.01	−.01	0	.06	.05
6	X	.52	.52	0	−.08	−.08	X	.55	.55	0	.02	.02	0	.05	.04
7	X	.76	.76	0	.05	.04	0	−.09	−.09	0	.06	.06	X	.55	.57
8	X	.44	.45	0	−.04	−.04	0	.24	.24	0	−.01	.00	0	.27	.25
9	X	.55	.55	X	.52	.52	0	−.08	−.08	0	.07	.06	0	.01	.02
10	X	.29	.29	X	.44	.45	X	.43	.44	X	.34	.31	0	−.03	−.01
11	X	.73	.73	0	−.16	−.15	0	−.17	−.17	0	.30	.29	0	−.09	−.08
12	X	.72	.72	0	−.01	−.02	0	.02	.02	0	.00	.00	X	.51	.50
13	X	.75	.75	0	−.01	.00	0	.06	.05	0	−.08	−.08	0	.13	.12
14	X	.50	.50	0	.01	.01	0	−.03	−.03	X	−.51	−.53	0	−.05	−.05
15	X	.56	.56	X	.54	.53	0	−.06	−.06	0	.04	.03	0	.01	.01
16	X	.34	.35	X	.34	.34	0	.28	.28	0	−.01	.00	0	.10	.09
17	X	.43	.43	0	−.08	−.09	X	.68	.68	0	−.05	−.04	0	.00	−.02
18	X	.56	.56	X	.56	.55	0	.16	.17	0	−.03	−.04	0	.09	.08
19	X	.65	.66	0	.01	.02	0	.19	.19	0	−.19	−.18	0	−.04	−.05
20	X	.75	.76	0	−.03	−.03	0	−.03	−.05	0	−.01	.01	0	−.15	−.15
21	X	.69	.69	0	.09	.09	0	.04	.03	0	−.15	−.14	0	−.08	−.08
22	X	.53	.52	X	.56	.57	0	−.05	−.05	0	−.04	−.05	0	−.05	−.04
23	X	.50	.51	X	.56	.56	0	−.12	−.12	0	−.04	−.03	0	.01	.01
24	X	.46	.47	X	.57	.57	0	−.13	−.13	0	−.11	−.11	0	−.07	−.07

Note. TP = Target Pattern, where X indicates unspecified elements whose magnitude is determined by rotation; MH = target rotated MH-RM estimates; EM = target rotated EM estimates.

sion is reduced to 5 in EM. Thus, the E-step needs a total of $5^5 = 3125$ function evaluations for the integral approximation. The EM algorithm required 1 hours 27 minutes of run time in 301 cycles. The MH-RM algorithm required 95 seconds of run time in 540 cycles. The EM algorithm used over 50 times more CPU time than MH-RM.

Table 7 compares the rotated factor loadings obtained from MH-RM and EM. Both solutions generally fit within the hypothesized structure and are close to each other. The overall RMSD from target is equal to .103 for MH-RM and .101 for EM. Note that the factor pattern presented here is just *one* possible solution and there is nothing irrevocable about the target pattern. The MH-RM solution has a log-likelihood of −18967.4, which is very close to the height of the log-likelihood (−18968.7) that the EM algorithm attained.

5.3. Independent Verifications

Mplus (Muthén & Muthén, 2008) is used to independently verify the solutions obtained by IRTPRO's EM module for the social quality of life data. Also of interest is the run time. Figure 1 shows a head-to-head comparison of the computing time of the three approaches (MH-RM in IRTPRO, EM in IRTPRO, and EM in Mplus), varying the numbers of factors extracted. For the unidimensional model, 21 quadrature points are used; for all other number of factors, there are 5 quadrature points per dimension. While the estimates as well as the log-likelihood values are virtually identical for all three programs, large difference in computing time surfaced for 4- and

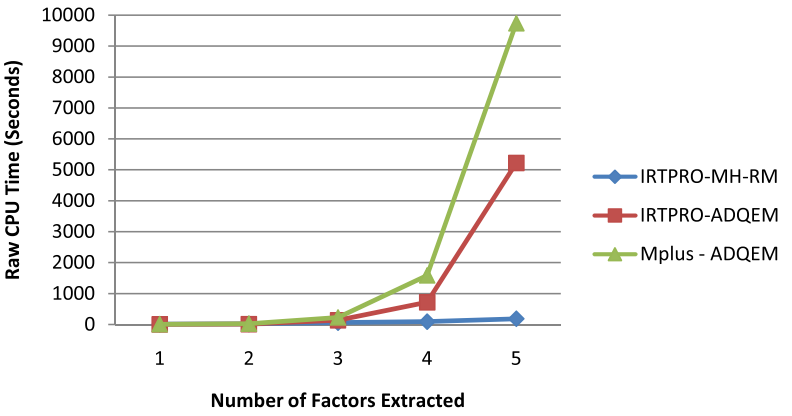

FIGURE 1.
Program run time comparisons. IRTPRO-MH-RM denotes the implementation of MH-RM algorithm in IRTPRO. IRTPRO-ADQEM denotes the implementation of the adaptive quadrature based EM algorithm in IRTPRO. Mplus-ADQEM denotes the implementation of the adaptive quadrature based EM algorithm in Mplus.

5-dimensional models. One can clearly see that the computing time for adaptive quadrature based EM is exponentially increasing while that of MH-RM grows only linearly as more factors are extracted. As witnessed in the simulation study in Section 4.4, for sufficiently low dimensional problems, MH-RM can be less efficient than EM, but in those situations, the absolute amount of time required is not large; whereas for high-dimensional problems, MH-RM can be far more efficient than EM because it does not involve numerical quadrature.

6. Extensions

6.1. Confirmatory Item Factor Analysis

Confirmatory item factor analysis is ideal when there exists sufficient prior theory about the factor structure of the items. The MCMC sampling procedure and the RM step in the proposed MH-RM algorithm remain the same as before. The constraints on the item parameters are imposed on the complete data log-likelihood. Estimation of the factor inter-correlations is straightforward because the complete data likelihood consists of two independent parts (see Equation (8)) so that the estimation of population distribution parameters is separate from the estimation of item parameters.

6.2. Explanatory and Multilevel IRT

Recent interest in generalizing the standard IRT model to include covariate effects, possibly even random-effects, has called for new estimation algorithms (de Boeck & Wilson, 2004; Fox & Glas, 2001; Fox, 2003, 2005). The MH-RM algorithm is uniquely suited to the goal of finding the MLEs for these extended IRT models because once the latent variables are "filled-in," the complete data model often takes the form of a generalized linear model. For instance, in explanatory IRT one can allow the item difficulty and slope parameters to depend on an observed covariate. The MH-RM solution to this problem simply involves conditioning on the covariate in addition to the observed response patterns when generating the factor score imputations, in which case the complete data model is specified as a generalized linear model with main effects for the covariate and the factor scores, as well as a term for their interaction. For multilevel IRT, the cluster-level random-effects induced by the random regression coefficients can also be thought

of as missing data that must be imputed along with the factor scores. In this case the complete data model usually reduces to a generalized linear model with an "offset" term (McCullagh & Nelder, 1989).

6.3. Other Types of Models

The MH-RM algorithm as implemented in its current form is designed with efficient item factor analysis as the primary objective. However, the algorithm can in principle be applied to fit more general kinds of latent variables models such as the comprehensive structural equation model proposed by Muthén (1984), the generalized linear latent variable model of Bartholomew and Knott (1999), and the GLLAMM model of Rabe-Hesketh, Skrondal, and Pickles (2004b). It is also worth noting that because item response theory models can be represented as nonlinear mixed-effects models (de Boeck & Wilson, 2004), the MH-RM algorithm may turn out to be a useful computational tool for parameter estimation in general nonlinear multilevel models. The SAEM algorithm, a close relative of the MH-RM algorithm (see Section 3.3), was initially intended for the mixture problem (Celeux & Diebolt, 1991). Thus, it is conceivable that the MH-RM may be extended to the case of categorical latent variables.

7. Discussion

This paper is concerned with the theoretical properties, implementation details, and empirical performance of a new parameter estimation algorithm for computing the MLE in the context of high-dimensional item factor analysis. The Metropolis–Hasting Robbins–Monro algorithm is a juxtaposition of elements from MCMC and stochastic approximation that appears to be more efficient than quadrature based EM for high-dimensional problems. For practical data analysis, the decrease in computational burden as witnessed in Section 5 translates into increase in productivity. Though the highest number of dimensions reported in the empirical applications in this paper is only five, Cai (2008a) already applied MH-RM successfully to problems with even higher dimensionality, where the speed difference is even more striking.

Before Bock and Aitkin (1981), the Bock and Lieberman (1970) style Newton algorithm simply could not handle the number of items that one encounters in day-to-day data analysis for testing situations. The EM algorithm due to Bock and Aitkin (1981) made IRT modeling practical for educational and psychological measurement. As IRT continues to evolve, high-dimensional IFA has become a valuable tool in at least some of its new domains of application, e.g., mental health and outcomes research. MH-RM provides a promising solution to the challenging numerical problems that arise from high-dimensional IFA. More research is still needed to study its behavior in a wide variety of situations and for models other than IFA.

Appendix A. The Convergence of MH-RM

The method for showing the convergence of the MH-RM algorithm relies on an Ordinary Differential Equation (ODE) argument that has become the predominant approach in the Stochastic Approximation literature (see Benveniste et al., 1990; Borkar, 2008; Kushner & Yin, 1997). It is instructive to illustrate this technique with the original Robbins–Monro algorithm. Equation (13) can be rewritten as a stochastic difference equation:

$$\theta_{k+1} = \theta_k + \gamma_k g(\theta_k) + \gamma_k \zeta_{k+1}, \tag{25}$$

where the last part $\gamma_k \zeta_{k+1}$ has mean zero and is independent from the past, so it can be thought of as a disturbance term. Due to the assumptions on the decaying gain constants, the asymptotic effect of the disturbance term can be regarded as negligible as k tends to infinity. Thus, when k is large and consequently γ_k is small, the remaining parts of (25) becomes

$$\theta_{k+1} = \theta_k + \gamma_k g(\theta_k). \tag{26}$$

Equation (26) can be considered a discrete approximation to the trajectory of the following ODE:

$$\frac{d\theta(t)}{dt} = g\big(\theta(t)\big), \tag{27}$$

similar to Euler's scheme for numerically integrating (27) given initial values $(\theta_0, g(\theta_0))$:

$$\theta_{k+1} = \theta_k + \gamma g(\theta_k). \tag{28}$$

As long as the ODE (27) is well posed, which is usually the case for realistic models and applications, the root-finding problem amounts to finding the equilibrium solution of the ODE. Using the foregoing argument, the Robbins–Monro recursions can be shown to asymptotically (in time) track the ODE with probability one (see, e.g., Borkar, 2008). Thus, the Robbins–Monro sequence should converge to a stable equilibrium of (27) and the root of $g(\theta) = 0$ can be found without even explicitly knowing the precise form of $g(\cdot)$.

Returning to the MH-RM algorithm, recall that $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. Reference to $\mathbf{Y}$ will be suppressed because it is fixed once observed. To avoid intricate notation, it is sufficient to consider $m_k = 1$ for all k. Let

$$\bar{\mathbf{H}}(\boldsymbol{\theta}) = \int_{\mathcal{E}} \mathbf{H}(\boldsymbol{\theta}|\mathbf{Z}) \Pi(d\mathbf{X}|\boldsymbol{\theta}), \quad \text{and} \quad \bar{\mathbf{s}}(\boldsymbol{\theta}) = \int_{\mathcal{E}} \mathbf{s}(\boldsymbol{\theta}|\mathbf{Z}) \Pi(d\mathbf{X}|\boldsymbol{\theta}).$$

Inspection of the stochastic difference equations (17) and (18) shows that the following set of ODEs govern the asymptotic (in time) behavior of MH-RM:

$$\begin{pmatrix} \frac{\partial}{\partial t}\boldsymbol{\theta}(t) \\ \frac{\partial}{\partial t}\boldsymbol{\Gamma}(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}(t)^{-1}\bar{\mathbf{s}}(\boldsymbol{\theta}(t)) \\ \bar{\mathbf{H}}(\boldsymbol{\theta}(t)) - \boldsymbol{\Gamma}(t) \end{pmatrix}, \qquad \begin{pmatrix} \boldsymbol{\theta}(0) \\ \boldsymbol{\Gamma}(0) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\Gamma} \end{pmatrix}. \tag{29}$$

It can be verified that for MLE $\hat{\boldsymbol{\theta}}$, the point $(\hat{\boldsymbol{\theta}}, \bar{\mathbf{H}}(\hat{\boldsymbol{\theta}}))$ is a stable equilibrium of the ODE (29). Therefore, the sequence of estimates generated by the MH-RM method converges to the MLE:

$$\boldsymbol{\theta}^{(k)} \to \hat{\boldsymbol{\theta}}, \quad \text{with probability 1 as } k \to \infty. \tag{30}$$

This result is a direct consequence of Gu and Kong (1998) Theorem 1, which is in turn based on general convergence results in Benveniste et al. (1990). The conditions needed for the convergence to hold are the same as those in Gu and Kong (1998) Theorem 1. These conditions guarantee (a) the integrability, convergence, and continuity of the Markov transition kernel, (b) the continuity and the existence of sufficient moments for functions $\mathbf{H}(\boldsymbol{\theta}|\mathbf{Z})$ and $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Z})$, and (c) that the process $\{(\boldsymbol{\theta}^{(k)}, \boldsymbol{\Gamma}_k), k \geq 1\}$ as defined by Equation (18) is a bounded sequence.

Appendix B. Complete Data Log-Likelihood and Derivatives

Suppressing references to i and j, the log-likelihood under consideration can be written using simplified notation as

$$l = \sum_{k=0}^{C-1} \chi_k(y) \log(P_k - P_{k+1}), \tag{31}$$

where $P_k = P(y \geq k|\boldsymbol{\theta}, \mathbf{x})$ is as defined in Equation (1). The first derivatives of (31) are

$$\frac{\partial l}{\partial \alpha_k} = -\left(\frac{\chi_{k-1}(y)}{P_{k-1} - P_k} - \frac{\chi_k(y)}{P_k - P_{k+1}}\right)\frac{\partial P_k}{\partial \alpha_k},$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{k=0}^{C-1} \frac{\chi_k(y)}{P_k - P_{k+1}}\left(\frac{\partial P_k}{\partial \boldsymbol{\beta}} - \frac{\partial P_{k+1}}{\partial \boldsymbol{\beta}}\right),$$

where $\partial P_k/\partial \alpha_k = P_k(1 - P_k)$, $\partial P_k/\partial \boldsymbol{\beta} = P_k(1 - P_k)\mathbf{x}$. The second derivatives are given by

$$\begin{aligned}
\frac{\partial^2 l}{\partial \alpha_k^2} &= -\left(\frac{\chi_{k-1}(y)}{(P_{k-1} - P_k)^2} + \frac{\chi_k(y)}{(P_k - P_{k+1})^2}\right)\left(\frac{\partial P_k}{\partial \alpha_k}\right)^2 \\
&\quad - \left(\frac{\chi_{k-1}(y)}{P_{k-1} - P_k} - \frac{\chi_k(y)}{P_k - P_{k+1}}\right)\left(\frac{\partial}{\partial \alpha_k}\frac{\partial P_k}{\partial \alpha_k}\right), \\
\frac{\partial^2 l}{\partial \gamma_{k-1}\partial \alpha_k} &= \frac{\chi_{k-1}(y)}{(P_{k-1} - P_k)^2}\left(\frac{\partial P_{k-1}}{\partial \gamma_{k-1}}\right)\left(\frac{\partial P_k}{\partial \alpha_k}\right), \\
\frac{\partial^2 l}{\partial \gamma_{k+1}\partial \alpha_k} &= \frac{\chi_k(y)}{(P_k - P_{k+1})^2}\left(\frac{\partial P_{k+1}}{\partial \gamma_{k+1}}\right)\left(\frac{\partial P_k}{\partial \alpha_k}\right), \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \alpha_k} &= -\frac{\chi_k(y)}{(P_k - P_{k+1})^2}\left(\frac{\partial P_k}{\partial \alpha_k}\right)\left(\frac{\partial P_k}{\partial \boldsymbol{\beta}} - \frac{\partial P_{k+1}}{\partial \boldsymbol{\beta}}\right) \\
&\quad + \frac{\chi_{k-1}(y)}{(P_{k-1} - P_k)^2}\left(\frac{\partial P_k}{\partial \alpha_k}\right)\left(\frac{\partial P_{k-1}}{\partial \boldsymbol{\beta}} - \frac{\partial P_k}{\partial \boldsymbol{\beta}}\right) \\
&\quad - \left(\frac{\chi_{k-1}(y)}{P_{k-1} - P_k} - \frac{\chi_k(y)}{P_k - P_{k+1}}\right)\left(\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial P_k}{\partial \alpha_k}\right), \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= \sum_{k=0}^{C-1}\left\{-\frac{\chi_k(y)}{(P_k - P_{k+1})^2}\left(\frac{\partial P_k}{\partial \boldsymbol{\beta}} - \frac{\partial P_{k+1}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial P_k}{\partial \boldsymbol{\beta}'} - \frac{\partial P_{k+1}}{\partial \boldsymbol{\beta}'}\right)\right. \\
&\quad \left. + \frac{\chi_k(y)}{P_k - P_{k+1}}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial P_k}{\partial \boldsymbol{\beta}'} - \frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial P_{k+1}}{\partial \boldsymbol{\beta}'}\right)\right\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial}{\partial \alpha_k}\frac{\partial P_k}{\partial \alpha_k} &= P_k(1 - P_k)(1 - 2P_k), \\
\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial P_k}{\partial \alpha_k} &= P_k(1 - P_k)(1 - 2P_k)\mathbf{x}, \\
\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial P_k}{\partial \boldsymbol{\beta}'} &= P_k(1 - P_k)(1 - 2P_k)\mathbf{x}\mathbf{x}'.
\end{aligned}$$

References

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

Aptech Systems, Inc. (2003). *GAUSS (Version 6.08) [Computer software]*. Maple Valley: Author.

Baker, F.B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York: Dekker.

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bartholomew, D.J., & Leung, S.O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1–15.
Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–561.
Benveniste, A., Métivier, M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*. Berlin: Springer.
Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge: MIT Press.
Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.
Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280.
Bock, R.D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.
Bolt, D. (2005). Limited and full information estimation of item response theory models. In A. Maydeu-Olivares & J.J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27–71). Mahwah: Earlbaum.
Booth, J.G., & Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society—Series B*, *61*, 265–285.
Borkar, V.S. (2008). *Stochastic approximation: A dynamical systems viewpoint*. Cambridge: Cambridge University Press.
Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.
Browne, M.W., Cudeck, R., Tateneni, K., & Mels, G. (2008). CEFA: Comprehensive Exploratory Factor Analysis (Version 3.02) [Computer software]. Retrieved from http://quantrm2.psy.ohio-state.edu/browne/.
Cai, L. (2006). *Full-information item factor analysis by Markov chain Monte Carlo stochastic approximation*. Unpublished master's thesis, Department of Statistics, University of North Carolina at Chapel Hill.
Cai, L. (2008a). *A Metropolis–Hastings Robbins–Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. Unpublished doctoral dissertation, Department of Psychology, University of North Carolina at Chapel Hill.
Cai, L. (2008b). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309–329.
Cai, L., du Toit, S.H.C., & Thissen, D. (2009, forthcoming). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Chicago: SSI International.
Cai, L., Maydeu-Olivares, A., Coffman, D.L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194.
Camilli, G. (1994). Origin of the scaling constant $d = 1.7$ in item response theory. *Journal of Educational and Behavioral Statistics*, *19*, 379–388.
Celeux, G., Chauveau, D., & Diebolt, J. (1995). *On stochastic versions of the EM algorithm* (Tech. Rep. No. 2514). The French National Institute for Research in Computer Science and Control.
Celeux, G., & Diebolt, J. (1991). *A stochastic approximation type EM algorithm for the mixture problem* (Tech. Rep. No. 1383). The French National Institute for Research in Computer Science and Control.
Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327–335.
de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.
Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, *27*, 94–128.
Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society—Series B*, *39*, 1–38.
Diebolt, J., & Ip, E.H.S. (1996). Stochastic EM: Method and application. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 259–273). London: Chapman and Hall.
Dunson, D.B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society—Series B*, *62*, 355–366.
Edwards, M.C. (2005). *A Markov chain Monte Carlo approach to confirmatory item factor analysis*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.
Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.
Fox, J.-P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology*, *56*, 65–81.
Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*, 145–172.
Fox, J.-P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 269–286.
Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.
Gu, M.G., & Kong, F.H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *The Proceedings of the National Academy of Sciences*, *95*, 7270–7274.
Gu, M.G., Sun, L., & Huang, C. (2004). A universal procedure for parametric frailty models. *Journal of Statistical Computation and Simulation*, *74*, 1–13.

Gu, M.G., & Zhu, H.-T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society—Series B*, *63*, 339–355.

Gueorguieva, R.V., & Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, *96*, 1102–1112.

Haberman, S.J. (1977). Log-linear models and frequency tables with small expected cell counts. *The Annals of Statistics*, *5*, 1148–1169.

Hastings, W.K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Huber, P., Ronchetti, E., & Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society—Series B*, *66*, 893–908.

Jank, W.S. (2004). Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM. *Computational Statistics and Data Analysis*, *48*, 685–701.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, *52*, 5066–5074.

Kass, R., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models. *Journal of the American Statistical Association*, *84*, 717–726.

Kuhn, E., & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, *49*, 1020–1038.

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Kushner, H.J., & Yin, G.G. (1997). *Stochastic approximation algorithms and applications*. New York: Springer.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society—Series B*, *57*, 425–437.

Liu, Q., & Pierce, D.A. (1994). A note on Gauss–Hermite quadrature. *Biometrika*, *81*, 624–629.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society—Series B*, *44*, 226–233.

Makowski, D., & Lavielle, M. (2006). Using SAEM to estimate parameters of models of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*, 45–60.

Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979). *Multivariate analysis*. San Diego: Academic Press.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using $g^2(\text{dif})$ to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55–64.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society—Series B*, *42*, 109–142.

McCullagh, P., & Nelder, J.A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

McCulloch, C.E., & Searle, S.R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

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.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthén, & Muthén (2008). *Mplus (Version 5.0) [Computer software]*. Los Angeles: Author.

Natarajan, R., & Kass, R.E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, *95*, 227–237.

Naylor, J.C., & Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society—Series C*, *31*, 214–225.

Orchard, T., & Woodbury, M.A. (1972). A missing information principle: Theory and application. In L.M. Lecam, J. Neyman, & E.L. Scott (Eds.), *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability* (pp. 697–715). Berkeley: University of California Press.

Patz, R.J., & Junker, B.W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Patz, R.J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004b). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004a). *GLLAMM manual* (U.C. Berkeley Division of Biostatistics Working Paper Series, 160).

Raudenbush, S.W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*, 141–157.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care*, *45*, S22–31.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*, 400–407.

Roberts, G.O., & Rosenthal, J.S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, *16*, 351–367.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 17*.

Savalei, V. (2006). Logistic approximation to the normal: The KL rationale. *Psychometrika*, *71*, 763–767.

Schilling, S., & Bock, R.D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Segall, D.O. (1998). *IFACT computer program Version 1.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation [Computer software]*. Seaside: Defense Manpower Data Center.

Shi, J.-Q., & Lee, S.-Y. (1998). Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *51*, 233–252.

Song, X.-Y., & Lee, S.-Y. (2005). A multivariate probit latent variable model for analyzing dichotomous responses. *Statistica Sinica*, *15*, 645–664.

te Marvelde, J., Glas, v.G.C., & van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, *66*, 5–34.

Thissen, D. (2003). *MULTILOG 7 user's guide*. Chicago: SSI International.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, *2*, 309–322.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, *22*, 1701–1762.

Tierney, L., & Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.

Titterington, D.M. (1984). Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society—Series B*, *46*, 257–267.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–202.

Wei, G.C.G., & Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, *85*, 699–704.

Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79.

Zhu, H.-T., & Lee, S.-Y. (2002). Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte-Carlo method. *Statistics and Computing*, *12*, 175–183.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (2003). *BILOG-MG3 user's guide*. Chicago: SSI International.

Manuscript Received: 14 SEP 2008
Final Version Received: 25 APR 2009
Published Online Date: 28 JUL 2009