University of California
Los Angeles

Estimation of a Ramsay-Curve Item Response Model By The Metropolis-Hastings Robbins-Monro Algorithm

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Scott Lee Monroe

2013

Abstract of the Thesis

Estimation of a Ramsay-Curve Item Response Model By The Metropolis-Hastings Robbins-Monro Algorithm

by

Scott Lee Monroe

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Ying Nian Wu, Chair

In Ramsay curve item response theory (RC-IRT, Woods & Thissen, 2006), the shape of the latent trait distribution is estimated simultaneously with the item parameters. In its original implementation, RC-IRT is estimated via Bock and Aitkin's (1981) EM algorithm, which yields maximum marginal likelihood estimates. This method, however, does not produce the parameter covariance matrix as an automatic byproduct upon convergence. In turn, researchers are limited in when they can employ RC-IRT, as the covariance matrix is needed for many statistical inference procedures. The present research remedies this problem by estimating the RC-IRT model parameters by the Metropolis-Hastings Robbins-Monro (MH-RM, Cai, 2010) algorithm. An attractive feature of MH-RM is that the structure of the algorithm makes estimation of the covariance matrix quite convenient. Additionally, MH-RM is ideally suited for multidimensional IRT, whereas EM is limited by the "curse of dimensionality." When RC-IRT is further generalized to multiple latent dimensions, MH-RM would appear to be the logical choice for estimation.

The thesis of Scott Lee Monroe is approved.

Li Cai

Rick Paik Schoenberg

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2013

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In unidimensional IRT, two typical assumptions are that the item response functions (IRFs) are logistic and that the latent trait, $g(\theta)$, is normally distributed. These default assumptions are supported by years of successful use in the analysis of educational and psychological data. However, only one of these assumptions is necessary since many different IRF-$g(\theta)$ combinations can produce the same joint distribution of item responses (Duncan & MacEachern, 2008). A major advantage of retaining the logistic form for IRFs is that doing so preserves interpretability of the item parameters estimates. At the same time, abandoning the normality assumption is attractive because it is easy to imagine scenarios where assuming normality of $g(\theta)$ is likely a misspecification of the model (see e.g., Woods & Thissen, 2006). For example, sampling of participants in psychological or educational studies may suggest a mixture of normals, based on treatment and control groups. As another example, the distribution of proficiency latent variables may be expected to be nonnormally distributed in certain subpopulations of interest (e.g., English language learners). Consequently, researchers have developed several methods to estimate or characterize the shape of the latent trait distribution. These methods include empirical histograms (Bock & Aitkin, 1981), normal mixtures (Mislevy, 1984), Ramsay curves (Woods & Thissen, 2006; Woods, 2006), and Davidian curves (Woods & Lin, 2008).

Despite the availability of these techniques, and their theoretical appeal, most analyses in practice fall back on the traditional assumption of normality. One reason for this incongruity is that the methods need further development and generalization. As a primary example, as well as motivation for this research, standard errors for the item parameters are not currently available when using Ramsay curve IRT (RC-IRT). Without standard errors or, more generally, the observed information matrix, researchers are limited in when they can use RC-IRT. For instance, standard errors are routinely used in test assembly as part of the item selection process. Also, the observed information matrix is needed for some limited-information goodness-of-fit testing (see, e.g., Cai, Maydeu-Olivares, Coffman, & Thissen, 2006) and differential item functioning analyses (Lord, 1980; Langer, 2008).

All model parameters of RC-IRT, in its original implementation, are estimated using Bock and Aitkin's (1981) EM algorithm. Decades after its development, the method still enjoys extensive use because of its stability and, often, its speed. However, the EM algorithm does not yield the observed information matrix upon convergence. Consequently, standard errors for RC-IRT item parameters are not currently available. There are methods designed to address this deficiency within the framework of EM (see, e.g., Louis, 1982; Cai, 2008). However, an alternative strategy, adopted here, is to choose another estimation method that is more amenable to estimation of standard errors.

This research uses the Metropolis-Hastings Robbins-Monro algorithm (MH-RM, Cai, 2010) to perform maximum marginal likelihood (MML) estimation for RC-IRT, and to obtain the observed information matrix upon convergence. As noted above, Bock and Aitkin (1981) EM does not preclude approximation of the observed information matrix. Nevertheless, MH-RM is preferred here because it is better-suited to accomodate further generalizations of RC-IRT. More specifically, when future research generalizes RC-IRT (or a similar methodology) to multidimensional latent traits, MH-RM would seem to be a logical and attractive choice for estimation. This is because the MH-RM algorithm is, in some sense, designed to address the "curse of dimensionality" that limits the feasibility of EM in multidimensional IRT.

As an aside, I do not advocate the use of RC-IRT, specifically. Rather, I view latent trait density estimation, generally, as theoretically appealing and practically useful. The characterization of the density, however, may be accomplished by numerous methods, all serving the same purpose.

The remainder of this paper is organized as follows. Chapter 2 presents Samejima's (1969) graded response model. This is followed by a review of Ramsay curves and RC-IRT in Chapter 3. Chapters 4-6 review and compare Bock and Aitkin (1981) EM and MH-RM (Cai, 2010). In Chapter 7, the details of RC-IRT implementation for MH-RM are provided. Then, Chapter 8 presents a simulation study examining the accuracy of point estimates and standard error estimates. Chapter 9 contains an empirical study. Finally, the paper concludes in Chapter 10 with directions for future research.

CHAPTER 2

A Graded Response Model for IRT

This section introduces notation for a logistic IRT model for graded responses following Samejima (1969).

2.1 Some Notation

Let there be $i = 1, 2, \ldots, N$ respondents, and $j = 1, 2, \ldots, n$ items. Let C_j be the number of response categories for item j. And, let $\mathrm{u}_{ij} = h$, $h \in \{0, 1, \ldots, C_j - 1\}$ be the item response from person i to item j. Then, $\mathbf{u}_i$ is a vector of item responses from person i, and $\mathbf{U}$ is an $N \times n$ matrix of response patterns, whose ith row is $\mathbf{u}_i'$.

For the jth item, let a_j be the item slope. Let $\mathbf{c}_j = (c_{j1}, \ldots, c_{j(C_j-1)})'$, be a $(C_j - 1) \times 1$ vector of intercepts for item j. Parameters for item j are collected in the $C_j \times 1$ vector $\boldsymbol{\beta}_j = (a_j, \mathbf{c}_j)$. Collecting all of the item parameters, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \ldots, \boldsymbol{\beta}_n')'$ is a $(\sum_{j=1}^{n} C_j) \times 1$ vector. For the ith person, let θ_i be the latent trait score, and let $\boldsymbol{\theta}$ be the $N \times 1$ vector of latent traits scores for all respondents. Often, the θ_i are assumed to follow a normal distribution. However, in RC-IRT, they are i.i.d. random variables governed by $\boldsymbol{\eta}$, a vector of Ramsay curve parameters. The length of $\boldsymbol{\eta}$ is $v = \mathit{degree} + \text{number of } \mathit{knots} - 1$. The terms *degree* and *knots* will be discussed in Section 3.1 below.

Lastly, the Ramsay Curve is described by a set of rectangular quadrature points, x_q, where there are $q = 1, 2, \ldots, Q$ points. The points are equally spaced every 0.1, and the range can be determined by the analyst (e.g., choosing a range from -6 to 6 results in 121 points).

2.2 Observed and Complete Data Likelihood

Conditional on an individual's latent trait score, θ_i, and the item parameters, $\boldsymbol{\beta}_j$, the conditional probability for response $u_{ij} = k$ is given by

$$\begin{aligned} \pi_{ijk} &= T(\mathrm{u}_{ij} = k|\boldsymbol{\beta}_j, \theta_i) \\ &= \frac{1}{1+\exp(-a_j\theta_i - c_{jk})} - \frac{1}{1+\exp(-a_j\theta_i - c_{j(k+1)})} \\ &= T^*(k) - T^*(k+1), \end{aligned} \tag{2.1}$$

where $T^*(k)$ is the conditional probability that a response is in category k or higher. Also, the following response probabilities are defined: $T^*(0) = 1$ and $T^*(C_j) = 0$. Based on Equation (2.1), the conditional distribution of u_{ij} is a multinomial with C_j cells and cell probabilities π_{ijk},

$$f(\mathrm{u}_{ij}|\boldsymbol{\beta}_j, \theta_i) = \prod_{k=0}^{C_j-1} \pi_{ijk}^{\chi_k(u_{ij})}, \tag{2.2}$$

where $\chi_k(u)$ is an indicator function defined as

$$\chi_k(u) = \begin{cases} 1, & \text{if } u = k, \\ 0, & \text{otherwise} \end{cases}. \tag{2.3}$$

Further, conditional on the latent trait, item responses are assumed independent (Lord & Novick, 1968). Consequently, the conditional density of $\mathbf{u}_i$ is simply the product over the n items:

$$f(\mathbf{u}_i|\boldsymbol{\beta},\theta_i)=\prod_{j=1}^{n} f(\mathrm{u}_{ij}|\boldsymbol{\beta}_j,\theta_i). \tag{2.4}$$

Again, for RC-IRT, the shape of the latent distribution depends on the RC parameters, $\boldsymbol{\eta}$. Therefore, the density for any θ_i depends on $\boldsymbol{\eta}$, and will hereafter be expressed as $g(\theta_i|\boldsymbol{\eta})$. Also, let $\boldsymbol{\omega}=(\boldsymbol{\beta},\boldsymbol{\eta})$ be the vector of all model parameters with length $d=(\sum_{j=1}^{n} C_j)+\nu$. Thus, for a person sampled from this potentially nonnormal distribution, the marginal density of $\mathbf{u}_i$ is:

$$f(\mathbf{u}_i|\boldsymbol{\omega})=\int\prod_{j=1}^{n} f(\mathrm{u}_{ij}|\boldsymbol{\beta}_j,\boldsymbol{\theta})g(\theta|\boldsymbol{\eta})d\theta. \tag{2.5}$$

Taking the product across persons, the observed data likelihood is

$$L(\boldsymbol{\omega}|\mathbf{U})=\prod_{i=1}^{N}\left[\int\prod_{j=1}^{n} f(\mathrm{u}_{ij}|\boldsymbol{\beta}_j,\boldsymbol{\theta})g(\theta|\boldsymbol{\eta})d\theta\right]. \tag{2.6}$$

If we treat the latent traits as missing data, then the complete data are $\boldsymbol{\theta}$, coupled with the observed item responses, $\mathbf{U}$. The complete data likelihood, then, is a product over every respondent's item responses and latent trait densities. In factored form, this complete data likelihood is

$$\begin{aligned} L(\boldsymbol{\omega}|\mathbf{U},\boldsymbol{\theta}) &= \prod_{i=1}^{N}\left[g(\theta_i|\boldsymbol{\eta})\prod_{j=1}^{n} f(\mathrm{u}_{ij}|\boldsymbol{\beta}_j,\theta_i)\right] \\ &= \left[\prod_{i=1}^{N} g(\theta_i|\boldsymbol{\eta})\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{n} f(\mathrm{u}_{ij}|\boldsymbol{\beta}_j,\theta_i)\right]. \end{aligned} \tag{2.7}$$

If we write the likelihoods on the right-hand side of (2.7) as functions of the parameters, and take the log of both sides, we obtain

$$\log L(\boldsymbol{\omega}|\mathbf{U},\boldsymbol{\theta}) = \log L(\boldsymbol{\eta}|\boldsymbol{\theta}) + \log L(\boldsymbol{\beta}|\mathbf{U},\boldsymbol{\theta}). \tag{2.8}$$

Equation (2.8) reveals that the complete data log likelihood can be understood as the sum of two independent parts: a log likelihood component for the RC parameters, $\boldsymbol{\eta}$, and a log likelihood component for the item parameters, $\boldsymbol{\beta}$. Moreover, the latter part corresponds to n ordinal logistic regressions, one for each item. This structure implies that during estimation, each of these sets of parameters can be updated separately.

CHAPTER 3

Ramsay Curve Item Response Theory

Given the structure of Equation (2.8), we seek a model that describes the form of the latent trait density given "observed" $\boldsymbol{\theta}$. Equation (2.8) also makes clear that such a model does not depend on item parameters or responses. One such model is based on Ramsay curves.

3.1 A Review of Ramsay Curves

What follows is a brief and general overview of Ramsay curves (RCs). The mathematical details are beyond the scope of this research, but interested readers should consult Ramsay (2000) and de Boor (2001) for a thorough treatment. Also, Woods and Thissen (2006) provides a nice introduction to RCs, particularly from a psychometric perspective.

Imagine we somehow observe $\boldsymbol{\theta}$, a collection of latent trait scores. Once observed, the objective is to characterize the density of $\boldsymbol{\theta}$. One natural approach is to construct a histogram, which is the motivation for the Empirical Histogram method discussed in Chapter 1. Another approach is to use an RC, which, unlike a histogram, will produce a smooth curve describing the distribution.

Basically, the shape of the RC density is found by connecting a set of curves known as B-splines. The range and potential flexibility of the RC is determined by the analyst, through three choices: the range for x_q, the *degree*, and the number of *knots*. First, the range of x_q simply defines the support of the density of

$g(\theta|\boldsymbol{\eta})$. A typical range for standardized latent traits is from -4.5 to 4.5 or -6 to 6. Second, the degree refers to the degree of the polynomial for each B-spline. Higher degrees can accommodate sharper curves. Third, the knots are where the B-splines connect to one another. Typically, the knots are evenly spaced across the range of support. A greater number of knots also allows more flexibility in the RC. The second and third choices (i.e., degree and number of knots) determine the number of elements in $\boldsymbol{\eta}$, the vector of RC parameters. As mentioned above, the length of $\boldsymbol{\eta}$ is $v = \textit{degree} + \text{number of } \textit{knots} - 1$.

Together, the three choices determine the structure of the $\mathbf{B}^*$ matrix (see Equation (11), Woods & Thissen, 2006). Assuming that $\mathbf{B}^*$ can be obtained, the height of the RC at each θ_i is given by

$$g(\theta_i|\boldsymbol{\eta}) = \frac{\exp\left[\mathbf{B}^*(\theta_i)\boldsymbol{\eta}\right]}{C}, \tag{3.1}$$

where

$$C = \sum_{q=1}^{Q} \exp\left[\mathbf{B}^*(x_q)\boldsymbol{\eta}\right] \tag{3.2}$$

is the normalization constant that ensures $g(\theta_i|\boldsymbol{\eta})$ integrates to 1 and is a proper density.

The likelihood for the RC is

$$L(\boldsymbol{\eta}|\boldsymbol{\theta}) = \prod_{i=1}^{N} g(\theta_i|\boldsymbol{\eta}), \tag{3.3}$$

which can be recognized as one of the factors of Equation (2.7). Once estimates of $\boldsymbol{\eta}$ are obtained, they can be used in (3.1) to find $g(\theta_i|\boldsymbol{\eta})$ for a particular respondent or to construct the entire RC. In summary, the RC method provides a smooth curve approximation for the distribution of “observed” $\boldsymbol{\theta}$.

3.2 Considerations for RC-IRT Estimation

Recall the complete data log likelihood in Equation (2.8), which is equal to the sum of two components. Given this structure, estimation for RC-IRT is conceptually straightforward. Again, the structure implies that the RC parameters can be dealt with separately from the item parameters during estimation. And conveniently, whatever estimation scheme is used for the item parameters can likewise be used for the RC parameters. Consequently, the basic structure of any estimation scheme for unidimensional IRT needs few changes to accommodate RCs.

Nevertheless, there are practical considerations for RC-IRT estimation that must be addressed. First among these is whether there is enough information in the data (i.e., $\boldsymbol{\theta}$) to estimate the RC parameters. For traditional IRT, assuming a normally distributed latent trait, little is lost by specifying the range of x_q as -4.5 to 4.5. This is because it is known a priori that there is little density outside this range. However, with nonnormal $g(\theta|\boldsymbol{\eta})$, we have no such a priori knowledge. Thus, a more reasonable range would be, for example, from -6 to 6. However, this expanded range presents us with a problem.

In practice, there may be some regions of the latent trait scale over which little or no information about the RC parameters is available. As a result, the corresponding spline coefficient may become empirically underidentified. And due to the dependencies among the RC parameters, this may cause a failure in estimation for all elements of $\boldsymbol{\eta}$. To safeguard against this possibility, Woods and Thissen (2006) implement a diffuse prior density for $\boldsymbol{\eta}$, and use Bayesian maximum a posteriori estimates in lieu of maximum likelihood estimates.

Unlike typical Bayesian applications, the prior in this case does not reflect a priori knowledge of the parameter values. Instead, it is employed to stabilize estimation. Woods and Thissen (2006) specify a ν-variate normal prior to compute the RC posterior, where ν is the number of RC parameters. The mean, $\boldsymbol{\mu}$, of the prior are those coefficients that would reproduce a normal density. The variance, Σ, is a diagonal matrix with all diagonal elements equal to some constant, ς. And since Σ is diagonal, all of the marginals of the prior are normals with variance equal to ς. Since the purpose of the prior is to stabilize estimation, ς should be as large as possible while still allowing successful estimation.

One other related consideration involves the analyst's choices for the degree and number of knots used in the RC. Although it may be tempting to choose large values (since the resulting RC is more flexible), there are two reasons to exercise restraint. First, for any given dataset, as ν increases, so does the possibility of estimation failure. This, of course, is related to the need for a prior, just discussed. Second, as ν increases, so does the possibility of overparameterization. To address this latter concern, multiple models should be fitted, with a range of values for the degree and number of knots. Then, standard model selection criteria may be applied, such as the Bayesian information criterion (BIC) or the Hannan-Quinn information criterion (HQIC).

CHAPTER 4

Two Approaches to Estimation for RC-IRT

Now that we have models for the item responses and the "observed" latent traits, we can turn our attention to estimation. In a maximum likelihood framework, we seek to find the estimate of $\boldsymbol{\omega}$ that maximizes the observed log likelihood, $l(\boldsymbol{\omega}|\mathbf{U})$ (where l denotes log likelihood). This research considers two alternatives for RC-IRT estimation: Bock and Aitkin (1981) EM (BA-EM) and MH-RM (Cai, 2010). Both approaches have been applied to standard unidimensional IRT models and are implemented in available software (see, e.g., IRTPRO Cai, Thissen, & du Toit, 2011). Further, BA-EM has been used extensively for RC-IRT (Woods & Thissen, 2006; Woods, 2007, 2008).

Before discussing the specifics of either algorithm, a major difference in approaches should be highlighted. Both EM and MH-RM exploit relationships between the observed data log likelihood, $l(\boldsymbol{\omega}|\mathbf{U})$, and complete data log likelihood $l(\boldsymbol{\omega}|\mathbf{U}, \boldsymbol{\theta})$, but do so in different ways. In EM, the MLE is found by iteratively maximizing the conditional expectation of $l(\boldsymbol{\omega}|\mathbf{U}, \boldsymbol{\theta})$ over $\Pi(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})$, where $\Pi(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})$ is the posterior predictive distribution of missing data. Convergence results (Wu, 1983) show that the successive estimates will result in a (local) maximizer of $l(\boldsymbol{\omega}|\mathbf{U})$.

MH-RM, on the other hand, is at its core a root-finding algorithm. Let $\dot{l}$ denote the first derivative of the log likelihood. Then, the equation for which we seek the root is

$$\dot{l}(\boldsymbol{\omega}|\mathbf{U}) = \frac{\partial l(\boldsymbol{\omega}|\mathbf{U})}{\partial \boldsymbol{\omega}} = \mathbf{0}, \tag{4.1}$$

found by setting $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$ equal to zero. The root of Equation (4.1) is the MLE. Also, due to the so-called Fisher's identity (Fisher, 1925),

$$\dot{l}(\boldsymbol{\omega}|\mathbf{U}) = \int \dot{l}(\boldsymbol{\omega}|\mathbf{U}, \boldsymbol{\theta})\Pi(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})d\boldsymbol{\theta}. \tag{4.2}$$

Finding the root of the right-hand side of Equation (4.2) results in a (local) maximizer of $l(\boldsymbol{\omega}|\mathbf{U})$. MH-RM obtains the solution iteratively by drawing imputations from the posterior predictive distribution $\boldsymbol{\Pi}(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})$ and stabilizes the noise introduced by the random draws with the Robbins-Monro (Robbins & Monro, 1951) stochastic approximation method.

CHAPTER 5

Bock-Aitkin EM for RC-IRT

Woods and Thissen (2006), in the original RC-IRT implementation, embedded estimation of the Ramsay curve within Bock and Aitkin (1981) EM. This section provides a brief overview of the algorithm, and the modifications needed to accommodate estimation of the Ramsay curve.

5.1 A Review of the Bock and Aitkin (1981) EM Algorithm

Notably, Bock and Aitkin (1981) EM (BA-EM) is quadrature-based. This feature dictates how values are computed, and requires equations slightly different in form than those presented in (2.5) through (2.8). The primary distinction is that while MH-RM takes a summation over examinees, BA-EM takes a summation over quadrature points. As an aside, this latter summation is what can limit the practicality of BA-EM in multidimensional IRT. The number of quadrature points grows exponentially with the dimensionality of the latent trait, regardless of sample size. In the literature, this phenomenon is sometimes called the "curse of dimensionality."

Very generally, the EM algorithm (Dempster, Laird, & Rubin, 1977) iteratively maximizes the expectation of $l(\boldsymbol{\omega}|\mathbf{U}, \boldsymbol{\theta})$ over $\Pi(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})$, where $\Pi(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})$ is the posterior predictive distribution of missing data. The procedure alternates between E-steps (for *expectation*) and M-steps (for *maximization*) until convergence. For BA-EM, the steps take the following forms.

For the E-step, given observed data and current parameter estimates, the conditional expectation of the missing data, $\boldsymbol{\theta}$, is found. For each item, this conditional expectation is collected in the so-called E-step tables. For the M-step, for each item, the E-step tables are treated as observed data and logit analysis is performed. The resulting item parameter estimates are used in the next E-step.

5.2 Modifications to BA-EM estimation for RC-IRT

Provided the infrastructure for constructing the RC is in place (as discussed in Chapter 3), the modifications needed to use BA-EM for RC-IRT are quite minimal. First, revise the E-step by using the current characterization of $g(\theta|\boldsymbol{\eta})$ to find the conditional expectation of $\boldsymbol{\theta}$. As before, fill in the E-step tables for each item. Second, estimate the proportion of respondents at each quadrature point, denoted $N(x_q)$, by summing across all E-step tables. Following Woods and Thissen (2006), at this point, the scale is identified by standardizing $N(x_q)$ to have a mean of 0 and variance of 1.

Finally, in the M-step, update the RC parameters. Given the structure of Equation (2.8), this update occurs independently of the item parameter updates. Since the set of $N(x_q)$ is akin to a collection of "observed" latent trait scores, the RC methodology in Section 3.1 can be used to find updated estimates of $\boldsymbol{\eta}$. The updated RC is used in the next E-step.

CHAPTER 6

A Review of the Metropolis-Hastings Robbins-Monro Algorithm

What follows is a broad outline of the MH-RM algorithm. For full details, see Cai (2010). As its name suggests, the MH-RM algorithm couples stochastic imputation via a Metropolis Hastings sampler (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) with a Robbins-Monro (Robbins & Monro, 1951) root-finding algorithm for noise-corrupted functions. Pieced together appropriately, these two methods complement one another to facilitate maximum likelihood estimation.

Recall Fisher's identity in Equation (4.2), which motivates the MH-RM algorithm. If we can find the expectation of $\dot{l}(\boldsymbol{\omega}|\mathbf{U},\boldsymbol{\theta})$ over $\Pi(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\omega})$, then we know the value of $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$, the gradient of the observed data log likelihood. And, if we can sample from $\Pi(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\omega})$, we can find the requisite expectation by Monte Carlo approximation. Fortunately, $\Pi(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\omega})$ is proportional to $L(\boldsymbol{\omega}|\mathbf{U},\boldsymbol{\theta})$, which allows us to construct an MH sampler. This is the MH part of MH-RM.

By laws of large numbers, we can approximate $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$ with arbitrary precision by increasing the number of MH imputations. However, with MH-RM, such a brute force approach is misguided. From one cycle to the next, the approximation to $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$ is only used to find the direction of the update for $\boldsymbol{\omega}$. For the sake of efficiency, a more sensible approach is to use a small number of imputations. Of course, this renders the sequence of approximations rather noisy. The RM method, though, was designed for such situations. Despite the noise, the RM method allows us to find the root of $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$, that is, the MLE. This is the RM part of MH-RM.

CHAPTER 7

An MH-RM Approach to RC-IRT

Having sketched an outline of the MH-RM algorithm, more details are needed to understand the implementation for RC-IRT. Some of these details are specific to the RC-IRT model (e.g., derivatives, selection of a sampler), while others are provided for context.

The sections that follow detail choices and results for derivatives, gain constants, starting values, parameter updates, and standard errors.

7.1 Complete Log Likelihood Derivatives

As mentioned above, $\dot{l}(\boldsymbol{\omega}|\mathbf{U}, \boldsymbol{\theta})$ is used to approximate $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$. Thus, first derivatives of the complete log likelihood are clearly needed. In addition, second derivatives are used in MH-RM for two purposes. Following the convention established earlier, let $\ddot{l} = \partial^2 l/(\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}')$ denote the matrix of second derivatives of the log likelihood. In MH-RM, $\ddot{l}(\boldsymbol{\omega}|\mathbf{U}, \boldsymbol{\theta})$ is used to compute a scaling factor for the parameter update which ideally speeds convergence. Also, it is used to approximate the observed information matrix and, by extension, estimates of standard errors. How this is accomplished is discussed below.

The derivatives for the graded response model are standard results (see, e.g., Baker & Kim, 2004, p. 213-217). Derivatives for the RC parameters are contained in Appendix A.

7.2 Specification of Gain Constants

Let $k = 1, 2, \ldots, \infty$ index the iteration for the MH-RM algorithm. The gain constants γ_k, for $k \geq 1$, scale the updates and serve to slowly average out the noise in the approximations to $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$. For this to occur, the γ_k need to slowly decrease to zero, which is ensured by the following conditions,

$$\gamma_k \in (0, 1], \quad \sum_{k=1}^{\infty} \gamma_k = \infty, \text{ and } \sum_{k=1}^{\infty} \gamma_k^2 < \infty. \tag{7.1}$$

If the γ_k decrease too quickly, then the estimates for $\boldsymbol{\omega}$ may stabilize prematurely, before the MLE is reached. Alternatively, if the γ_k decrease too slowly, the estimates for $\boldsymbol{\omega}$ may never stabilize. One satisfactory option, noted by Cai (2010), is to take γ_k as $1/k$. The rate of the decrease can be further fine-tuned by taking γ_k as $1/k^{\epsilon}$, $1/2 < \epsilon \leq 1$ (Polyak & Juditsky, 1992).

7.3 Computing Updates

What follows is a presentation of Equations (16) through (18) from Cai (2010). There is nothing particular to RC-IRT that needs to be addressed here. Nevertheless, the material is included because it is essential to understanding other aspects of the RC-IRT implementation.

Recall that MH-RM seeks to find the root of $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$ by iteratively estimating the expectation of $\dot{l}(\boldsymbol{\omega}|\mathbf{U}, \boldsymbol{\theta})$ over $\Pi(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})$, and updating $\boldsymbol{\omega}$ accordingly. Provided appropriate gain constants are specified, and samples from $\Pi(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})$ can be obtained, the parameters are updated in the following manner. Let $d = (\sum_{j=1}^{n} C_j) + \nu$ be the number of parameters in the model. Then, let $(\boldsymbol{\omega}^{(0)}, \boldsymbol{\Gamma}_0)$ be initial values, where $\boldsymbol{\Gamma}_0$ is a $d \times d$ symmetric positive definite matrix. Let $\boldsymbol{\omega}^{(k)}$ be the parameter estimate at the end of iteration k. The $(k + 1)$th iteration consists of *stochastic imputation*, *stochastic approximation*, and an *RM update*.

For *stochastic imputation*, m_k sets of missing data $\{\boldsymbol{\theta}_p^{(k+1)}; p = 1, 2, \ldots, m_k\}$ are drawn from $\Pi(\boldsymbol{\theta}|\mathbf{U}, \boldsymbol{\omega})$, to form m_k complete data sets $\{(\mathbf{U}, \boldsymbol{\theta}_p^{(k+1)}); p = 1, 2, \ldots, m_k\}$. Again, the RM method renders large m_k unnecessary.

For *stochastic approximation*, using Fisher's (1925) identity, we approximate the observed data gradient, $\dot{l}(\boldsymbol{\omega}|\mathbf{U})$, by the sample average of complete data gradients,

$$\tilde{\mathbf{s}}^{(k+1)} = \frac{1}{m_k} \sum_{p=1}^{m_k} \dot{l}(\boldsymbol{\omega}^{(k)}|\mathbf{U}, \boldsymbol{\theta}_p^{(k+1)}). \tag{7.2}$$

Also, to speed convergence, we use curvature information by recursively approximating the conditional expectation of the complete data information matrix,

$$\boldsymbol{\Gamma}^{(k+1)} = \boldsymbol{\Gamma}_k + \gamma_k \left\{ -\frac{1}{m_k} \sum_{p=1}^{m_k} \ddot{l}(\boldsymbol{\omega}^{(k)}|\mathbf{U}, \boldsymbol{\theta}_p^{(k+1)}) - \boldsymbol{\Gamma}_k \right\}. \tag{7.3}$$

Finally, in the *RM update*, we set the new parameter estimate to

$$\boldsymbol{\omega}_{k+1} = \boldsymbol{\omega}_k + \gamma_k (\boldsymbol{\Gamma}_{k+1}^{-1} \tilde{\mathbf{s}}_{k+1}). \tag{7.4}$$

For RC-IRT, the above steps are carried out $n + 1$ times per iteration, once per item and once for the set of RC parameters. Again, this is a consequence of the independence implied by Equation (2.8).

7.4 Constructing an MH Sampler

As in Patz and Junker (1999) and Cai (2010), a Metropolis-within-Gibbs sampling scheme is used to impute the latent trait scores. Let $q(\theta_i, \theta_i^*)$ be the transition density for moving from θ_i to θ_i^*. Also, define the acceptance factor as

$$\alpha(\theta_i, \theta_i^*) = \min\left[\frac{f(\mathbf{u}_i|\boldsymbol{\beta}, \theta_i^*)g(\theta_i^*|\boldsymbol{\eta})q(\theta_i^*, \theta_i)}{f(\mathbf{u}_i|\boldsymbol{\beta}, \theta_i)g(\theta_i|\boldsymbol{\eta})q(\theta_i, \theta_i^*)}, 1\right]. \tag{7.5}$$

Notably, (7.5) depends on the RC parameters to calculate $g(\theta_i|\boldsymbol{\eta})$. However, the acceptance probability for each θ_i depends on neither the latent trait scores of other persons, nor their item responses. Thus, all draws can be performed simultaneously with vector operations. Commonly, a symmetric random walk chain with some scalar dispersion parameter is used (Metropolis et al., 1953) for the transition density. In such a case, $q(\theta_i^*, \theta_i) = q(\theta_i, \theta_i^*)$, and the acceptance factor can be further simplified.

However, for RC-IRT, a symmetric random walk chain proves problematic. The minimum and maximum for x_q are user-defined, but in any case the RC has no density outside this range. Consequently, for all $\theta_i^* \notin [x_{\min}, x_{\max}]$, $g(\theta_i^*|\boldsymbol{\eta}) = 0$, which implies $\alpha(\theta_i, \theta_i^*) = 0$. Clearly, any transition density routinely imputing values that cannot be accepted is inefficient. To address this issue, the implemented transition density is constructed so that $\alpha(\theta_i, \theta_i^*) \neq 0$ (except very rarely), regardless of the scalar dispersion parameter.

To accomplish this goal, we seek a transition density where $\text{Var}(\boldsymbol{\theta}^*) = 1$. Such a condition will ensure that the imputations only rarely fall outside of the range for x_q. Let δ be the scalar dispersion parameter, let $e_i \sim \mathcal{N}(0, 1)$, and let $\mathbf{e}$ be a vector of normal deviates whose ith element is e_i. As a reminder, $\text{Var}(\boldsymbol{\theta}) = 1$ to identify the scale. Then, let $\varphi = \text{Var}(\boldsymbol{\theta} + \delta e) = 1 + \delta^2$. Finally, let the proposal draws be generated by $\boldsymbol{\theta}^* = (\boldsymbol{\theta}/\sqrt{\varphi}) + (\delta/\sqrt{\varphi})\mathbf{e}$. It can easily be verified that $\text{Var}(\boldsymbol{\theta}^*) = 1$, which achieves the stated goal. For the simulation study and empirical example, $\delta = 2.4$ was used.

Written as the density function of a normal distribution, the transition density is

$$q(\theta_i, \theta_i^*) = \frac{1}{(\delta/\varphi)\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\frac{(\theta^* - (\theta/\sqrt{\varphi}))^2}{(\delta^2/\varphi)}\right\}. \tag{7.6}$$

Examining the exponent of Equation (7.6), it is clear that $q(\theta_i^*, \theta_i) \neq q(\theta_i, \theta_i^*)$. Consequently, the ratio of transition densities in Equation (7.5) cannot be further simplified.

7.5 Providing Reasonable Starting Values

Without reasonable starting values, MH-RM estimation may fail, particularly if there is little information for some parameters. This is unsurprising as the convergence theory for the algorithm depends on the use of sufficiently good starting values (Borkar, 2008). Consequently, starting values for MH-RM are quite important. Luckily, the flexibility of the method allows a clever solution to this problem.

To explain this solution, it helps to consider the algorithm as proceeding through three successive stages: Stage 1, Stage 2, and Stage 3. These will be explained momentarily. Similarly, it is useful to introduce two types of starting values: "crude" and "refined." Let us assume for a moment that if refined values are available to start Stage 3, then estimation will succeed. The goal, then, is to find refined values to start Stage 3.

This goal is achieved in the following way. First, crude values are provided to start Stage 1. For instance, set $a = 1$ for all slopes and $c = 0$ for all intercepts. These values are crude in that they are in no way based on the data. Next, we set γ_k and m_k equal to unity for all of the iterations in Stage 1. As noted in Cai (2010), the MH-RM algorithm, specified in this way, is a close relative to Diebolt and Ip's (1996) stochastic EM (SEM) algorithm. Importantly, the SEM-type iterations move $\hat{\boldsymbol{\omega}}_k$ quickly to the neighborhood of the MLE. Stage 1 should run as long as necessary for the analyst to be confident that $\hat{\boldsymbol{\omega}}_k$ has reached this neighborhood. This concludes Stage 1.

In Stage 2, the SEM-type iterations continue where Stage 1 left off, but with a different purpose. As a reminder, the goal of this process is to provide refined starting values for Stage 3. These refined values can be obtained by averaging $\hat{\boldsymbol{\omega}}_k$ for some number of Stage 2 iterations (e.g., 100). Upon obtaining these averages, Stage 2 is complete.

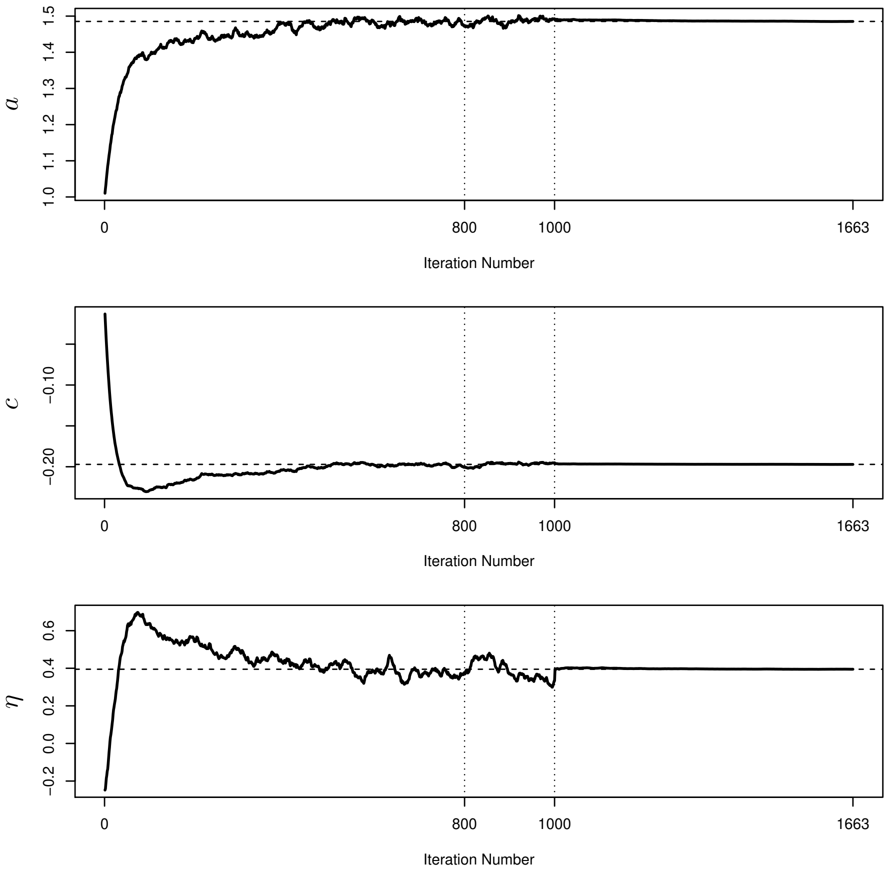
Finally, the refined values are used to start Stage 3. In this last stage, the SEM-type iterations are abandoned for iterations with decreasing gain constants. In this way, the noise is averaged out and the MLE is reached. This strategy is effective because the mean of the invariant distribution in Stage 2 is close to the MLE. As a result, there is a bit of leeway in the choice for the series of decreasing gain constants in Stage 3.

Figure 7.1 shows the Stages for three typical sequences of estimates from one replication of the simulation study (presented in Chapter 8). From the top, the three panels show the estimates for a slope parameter, an intercept parameter, and an RC parameter. In Stage 1 (SEM-type iterations $1 - 800$), the estimates move (relatively) quickly to reach the neighborhood of the MLE. In Stage 2 (SEM-type iterations $801 - 1000$), sample estimates are collected to compute means with which to start Stage 3. And lastly, in Stage 3 (decreasing γ_k iterations $1,001 - 1,663$), the estimates converge to the MLE.

7.6 Approximating the Observed Information

Louis (1982), derived a useful equality, linking the observed information to functions of the complete log likelihood. The information matrix of the observed log likelihood is

Figure 7.1: Example of MH-RM for RC-IRT: Sequences of estimates for three parameters



Note. a = slope parameter; c = intercept parameter; η = RC parameter. Horizontal lines (dashed) indicate MLE values. Vertical lines (dotted) demarcate the 3 stages of iterations. Stage 1 consisted of the first 800 iterations. Stage 2 consisted of the next 200 iterations. Stage 3 continued until the convergence criteria was reached at 1,663 iterations.

$$-\ddot{l}(\boldsymbol{\omega}|\mathbf{U}) = E_{\boldsymbol{\omega}}\{-\ddot{l}(\boldsymbol{\omega}|\mathbf{U},\boldsymbol{\theta})\} - \mathrm{cov}_{\boldsymbol{\omega}}\{\dot{l}(\boldsymbol{\omega}|\mathbf{U},\boldsymbol{\theta})\}, \tag{7.7}$$

where the expectation is with respect to $\Pi(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\omega})$. Cai (2010) proposed a method that uses Louis's equation, where the elements needed for computation are simply byproducts of the MH-RM estimation procedure (Cai, 2010, p. 42). This is one of the (two) methods of approximation implemented in IRTPRO (Cai et al., 2011). A benefit of this approach is that the observed information approximation is performed concurrently with parameter estimation.

Another approach, again following Louis (1982), is proposed by Diebolt and Ip (1996). Basically, the strategy uses Monte Carlo integration to approximate the mean and covariance in Equation (7.7). The parameter estimate, $\hat{\boldsymbol{\omega}}$, is fixed at the MLE, and P Monte Carlo samples of $\boldsymbol{\theta}$ are generated from $\Pi(\boldsymbol{\theta}|\mathbf{U},\boldsymbol{\omega})$. These samples are used to approximate the terms on the right-hand side of Equation (7.7). For some examples of this latter method, see Diebolt and Ip (1996) and Fox (2003). This method is also available in IRTPRO (Cai et al., 2011).

One last feature of the complete information matrix should be noted. Due to the independence of the logistic regression models, as well as the independence of the RC model, the complete data information matrix is block diagonal. Each item block is $C_j \times C_j$, and the RC block is $\nu \times \nu$. Hence, the entire matrix is $d \times d$, where $d = (\sum_{j=1}^{n} C_j) + \nu$.

CHAPTER 8

Simulation Study

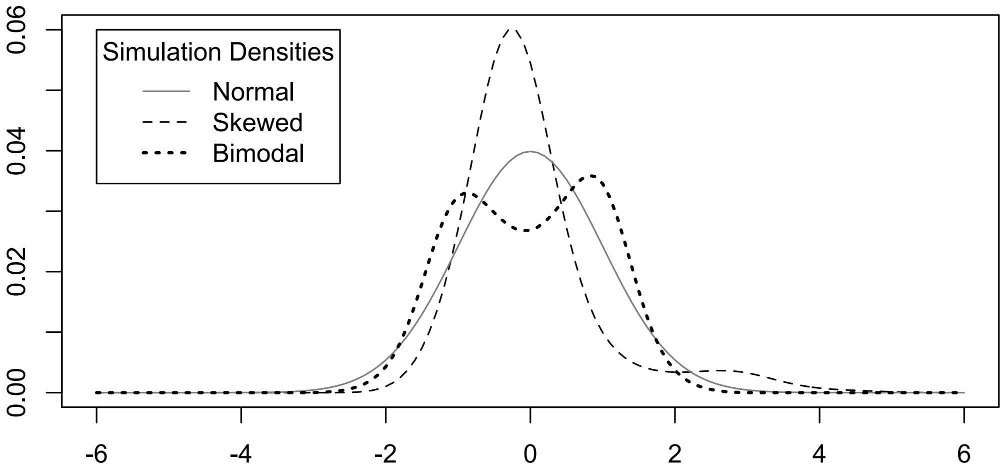
A Monte Carlo simulation study was conducted to compare the MH-RM and EM results, and to evaluate the accuracy of the MH-RM standard errors. The purpose of comparing MH-RM and EM is simply to verify that the MH-RM implementation is correct. As both methods compute MLEs, substantially discrepant results would indicate an improper implementation.

8.1 Methods and Design

This section details how the data were generated and how estimation was specified. Generally, data were generated to be realistic for psychological or educational research. For the study, there were $N = 1000$ simulees and $n = 25$ items. There were three conditions, based on the true shape of the latent trait distribution. Finally, there were 100 replications for each condition.

The true shape of $g(\theta|\boldsymbol{\eta})$ was either normal, skewed, or bimodal. All densities were represented by rectangular quadrature points, ranging from -6 to 6 by 0.1. The RC parameters for these densities were generated by mixing two normals. For the skewed density, the generating parameters were: $\mu_1 = -0.25$, $\mu_2 = 2.19$, $\sigma_1^2 = 0.37$, $\sigma_2^2 = 1.10$, $mp_1 = 0.9$, and $mp_2 = 0.1$. For the bimodal density, the values were: $\mu_1 = -1$, $\mu_2 = 1$, $\sigma_1^2 = 0.49$, $\sigma_2^2 = 0.49$, $mp_1 = 0.48$, and $mp_2 = 0.52$. The mixtures were then standardized to have $\mu = 0$ and $\sigma^2 = 1$. Next, the standardized mixtures were treated as data in RC likelihood functions

Figure 8.1: True densities used for the simulation study



Note. True normal curve (gray solid line); true skewed curve (dashed line); true bimodal curve (thick dotted line).

(see Equation (3.3)) with degree=5 and knots=6. Finally, the likelihoods were maximized to yield estimates of $\boldsymbol{\eta}$. These estimates were subsequently treated as the true RC parameters. The resulting densities can be seen in Figure 8.1.

The true item parameters were generated in a manner similar to Woods and Lin (2008). The slope parameters, a, were drawn from a truncated normal distribution with mean $=$ 1.8 and standard deviation $=$ 0.8, and truncation at 0.5 and 4. Difficulty parameters (typically labelled b) were drawn in the following way. The first difficulty parameter (b_1) was drawn from a normal with mean$=$ 1 and SD$=$ 0.5. To obtain b_2, a random draw was taken from a normal (mean$=$ 1 and SD$=$ 0.2) and *added* to b_1. Both b_3 and b_4 were drawn under the same procedure. Since the graded response model in Equation (2.1) is written in terms of slopes and intercepts, the latter were calculated as $c = -ab$. The same item parameters were used for all replications across all conditions, and are displayed in Table 8.1.

For each replication, $\boldsymbol{\theta}$ was drawn from the true $g(\theta|\boldsymbol{\eta})$ using rejection sampling (von Neumann, 1951). Then, probabilities of $u_{ij} = 1$ (i.e., correct item responses) were simulated according to the graded model in Equation (2.1). These proba-

Table 8.1: Simulation study: generating item parameter values

Item	Slope	Intercept 1	Intercept 2	Intercept 3	Intercept 4
1	1.89	1.71	−0.32	−2.59	−4.56
2	1.44	2.37	0.88	−0.66	−1.90
3	1.45	3.71	2.50	1.42	0.24
4	2.18	3.20	0.60	−1.24	−2.97
5	1.56	0.47	−1.66	−3.36	−5.11
6	1.81	3.28	1.42	−0.25	−2.69
7	0.50	1.07	0.58	0.20	−0.27
8	0.89	0.37	−0.48	−1.56	−2.66
9	0.94	1.00	−0.22	−1.24	−2.37
10	1.30	1.09	−0.32	−1.30	−2.60
11	1.56	2.05	0.20	−1.17	−3.17
12	1.37	2.28	0.51	−1.05	−1.82
13	3.18	5.04	2.11	−2.37	−5.47
14	2.02	3.15	1.05	−1.16	−2.65
15	2.34	3.93	1.97	0.44	−1.23
16	3.87	2.94	−1.16	−5.59	−10.14
17	1.63	1.03	−1.06	−2.70	−4.46
18	1.41	1.39	−0.31	−1.33	−2.93
19	0.75	0.62	−0.19	−1.13	−2.22
20	1.99	1.35	−0.37	−2.31	−3.90
21	1.99	−0.48	−2.54	−4.23	−6.34
22	1.98	0.04	−1.79	−3.29	−5.88
23	1.85	1.67	−0.52	−2.60	−4.14
24	1.57	0.35	−1.02	−2.23	−4.22
25	1.96	1.06	−0.86	−2.99	−5.14

bilities were compared to random uniform (0, 1) variables. Item responses were determined in proportion to the model-implied probabilities.

The graded model was fitted to the data in each replication, using both BA-EM and MH-RM. The starting values for all $a = 1$. For all $\mathbf{c}_j$, the starting values were $1, 1/3, -1/3$, and -1. The starting values for $\boldsymbol{\eta}$ were those that would reproduce a normal density. The density, $g(\theta|\boldsymbol{\eta})$, was represented using 121 rectangular quadrature points, evenly spaced from -6 to 6. Lastly, the RC was estimated using degree=5 and knots=6. Thus, for both the item responses and RC, the fitted and data generating models were the same.

As mentioned earlier, the forms of the RC likelihood functions for BA-EM and MH-RM take different forms and imply different scales. Thus, different values for ς (the variance for each marginal of the multivariate normal prior) are appropriate depending on the method of estimation. Based on trial and error, ς was set to 1 for MH-RM. For BA-EM, ς was set to 1000.

For MH-RM, the simulation size m_k was set to 1 for all iterations. Like in the example in Section 7.5, Stage 1 consisted of 800 iterations, followed by Stage 2 with 200 iterations. For Stage 3, the decreasing gain constants were set to $\gamma_k = 1/k^{\epsilon}$ with $\epsilon = 0.75$. For convergence criteria, the iterations were examined across a window of 3 iterations. Once the maximum absolute change across the window dropped below 1.0×10^{-4}, the iterations were deemed converged. For BA-EM, the iterations were considered converged once the maximum absolute between-iteration change in parameter estimates dropped below 1.0×10^{-4}.

8.2 Outcome Measurements

Overall model fit was assessed using log-likelihood values. Greater values indicate better fit. To evaluate the accuracy of item parameter recovery, estimated bias and root mean square error (RMSE) are used. Let M be the number of Monte Carlo replications and ω denote the true value of an arbitrary element of the parameter vector $\boldsymbol{\omega}$. Then, estimated bias is defined as $M^{-1}\sum_{m=1}^{M}(\hat{\omega}_m - \omega)$ where $\hat{\omega}_m$ is the MLE for ω in replication m. RMSE is defined as $\sqrt{M^{-1}\sum_{m=1}^{M}(\hat{\omega}_m - \omega)^2}$.

Since the scales of the true RC parameters are both unfamiliar and quite variable, estimated bias and RMSE are less appropriate measures of recovery accuracy. Instead, the integrated square error (ISE),

$$ISE(\hat{g}) = \int \{g(\theta|\hat{\boldsymbol{\eta}}) - g(\theta|\boldsymbol{\eta})\}^2 d\theta, \tag{8.1}$$

is used to measure the similarity between the true and estimated RCs, as in Woods and Lin (2008). The ISE was multiplied by 1,000 to facilitate comparison. Also, when aggregating the ISE statistic, the median instead of the mean was used due to skewness and kurtosis.

To assess the accuracy of the estimated standard errors, let $se(\hat{\omega})_m$ be the estimated standard error for ω in replication m. Then, $\overline{se}(\hat{\omega}) = M^{-1}\sum_{m=1}^{M} se(\hat{\omega})_m$. Also, let $sd(\hat{\omega}) = \sqrt{(M-1)^{-1}\sum_{m=1}^{M}(\hat{\omega}_m - \overline{\omega})^2}$ be the Monte Carlo standard deviation, where $\overline{\omega}$ is the mean of the estimates across replications. If the standard errors are estimated accurately, the averages, $\overline{se}(\hat{\omega})$, should correspond to the Monte Carlo standard deviations, $sd(\hat{\omega})$.

8.3 Results: Points Estimates from MH-RM and EM

All replications converged for both the MH-RM and BA-EM algorithms. For both algorithms, the log-likelihood (plus 30,000), ISE (multiplied by 1,000), and estimated bias and RMSE of item parameters are displayed in Table 8.2. Generally, the comparability of BA-EM and MH-RM is established by comparing means of outcome measurements across replications, as well as inspecting plots of these measurements by replicaton.

Table 8.2: Simulation Results for MH-RM and BA-EM Estimations of RC-IRT

	LogL	RMSE: a	Bias: a	RMSE: c	Bias: c	ISE
	Normal $g(\theta\|\boldsymbol{\eta})$					
MH-RM	−500.10	0.10	0.01	0.14	−0.01	0.04
BA-EM	−500.11	0.10	0.01	0.14	0.02	0.04
	Skewed $g(\theta\|\boldsymbol{\eta})$					
MH-RM	−393.75	0.12	0.01	0.13	0.00	0.09
BA-EM	−393.07	0.12	0.01	0.13	0.00	0.09
	Bimodal $g(\theta\|\boldsymbol{\eta})$					
MH-RM	−381.78	0.10	0.01	0.14	−0.00	0.13
BA-EM	−381.77	0.09	−0.00	0.14	0.01	0.14

Note. LogL = log-likelihood (plus 30,000); RMSE = root mean square error; Bias = Monte Carlo average estimate minus the true parameter value; a = slope parameter; c = intercept parameter; ISE = median of the integrated square error multiplied by 1,000.

For each $g(\theta|\boldsymbol{\eta})$ shape, the average log-likelihoods for MH-RM and BA-EM estimations are virtually identical. In addition to the average log-likelihoods, the values at each replication are extremely similar. In Figure 8.2, all of the points are very close to the 45° reference line, regardless of $g(\theta|\boldsymbol{\eta})$. Thus, in terms of global fit, there is no appreciable difference between the MH-RM and BA-EM results.

Again consulting Table 8.2, the average RMSE and estimated bias for the item parameters are nearly indistinguishable under MH-RM and BA-EM. Assessing item parameter recovery from another perspective, Figure 8.3 compares the average RMSE for all item parameters, by replication, for MH-RM and BA-EM. Again, the vast majority of points are quite close to the 45° reference line, indicating that for each replication MH-RM and BA-EM are producing comparable results.

Results for the accuracy of the estimated RCs is also presented in Table 8.2 via the ISE statistic. As should be expected, the ISE values for the normal $g(\theta|\boldsymbol{\eta})$ are

Figure 8.2: Log-likelihood Values for MH-RM (x-axis) and BA-EM (y-axis) Algorithms

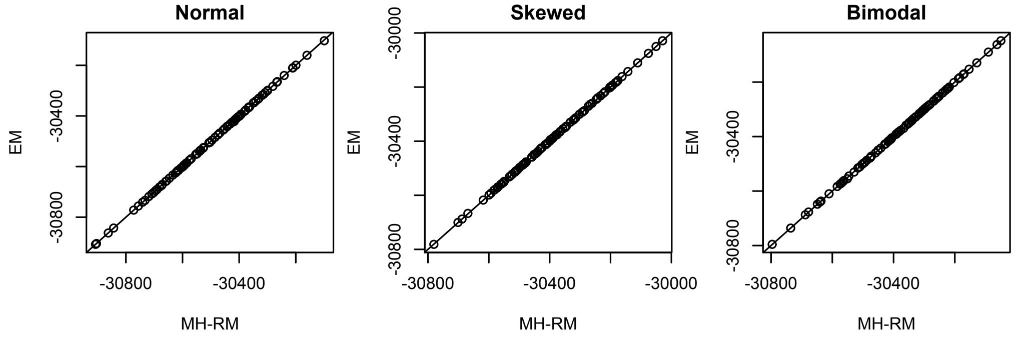


Figure 8.3: Average RMSE for Item Parameters within a Replication for MH-RM (x-axis) and BA-EM (y-axis) Algorithms

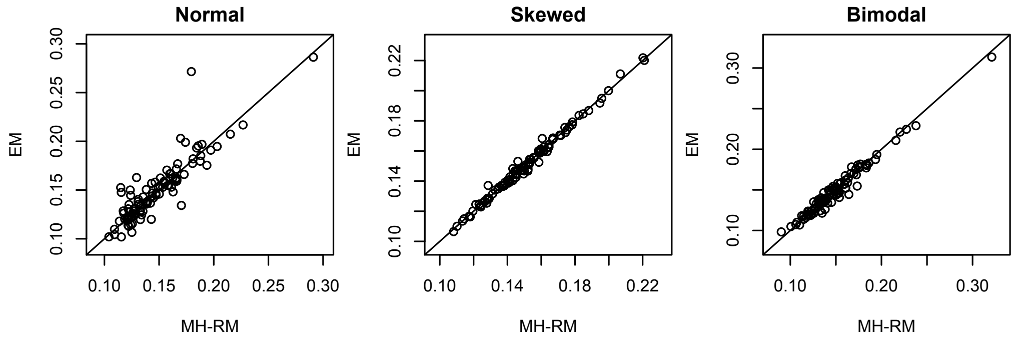
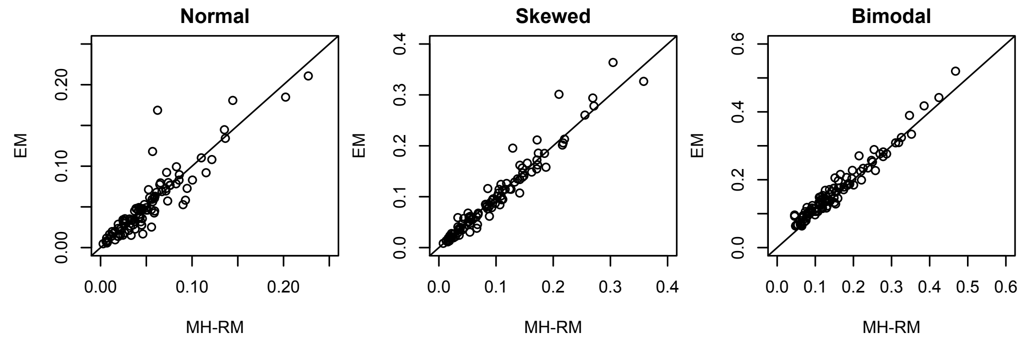


Figure 8.4: ISE for MH-RM (x-axis) and BA-EM (y-axis) Algorithms



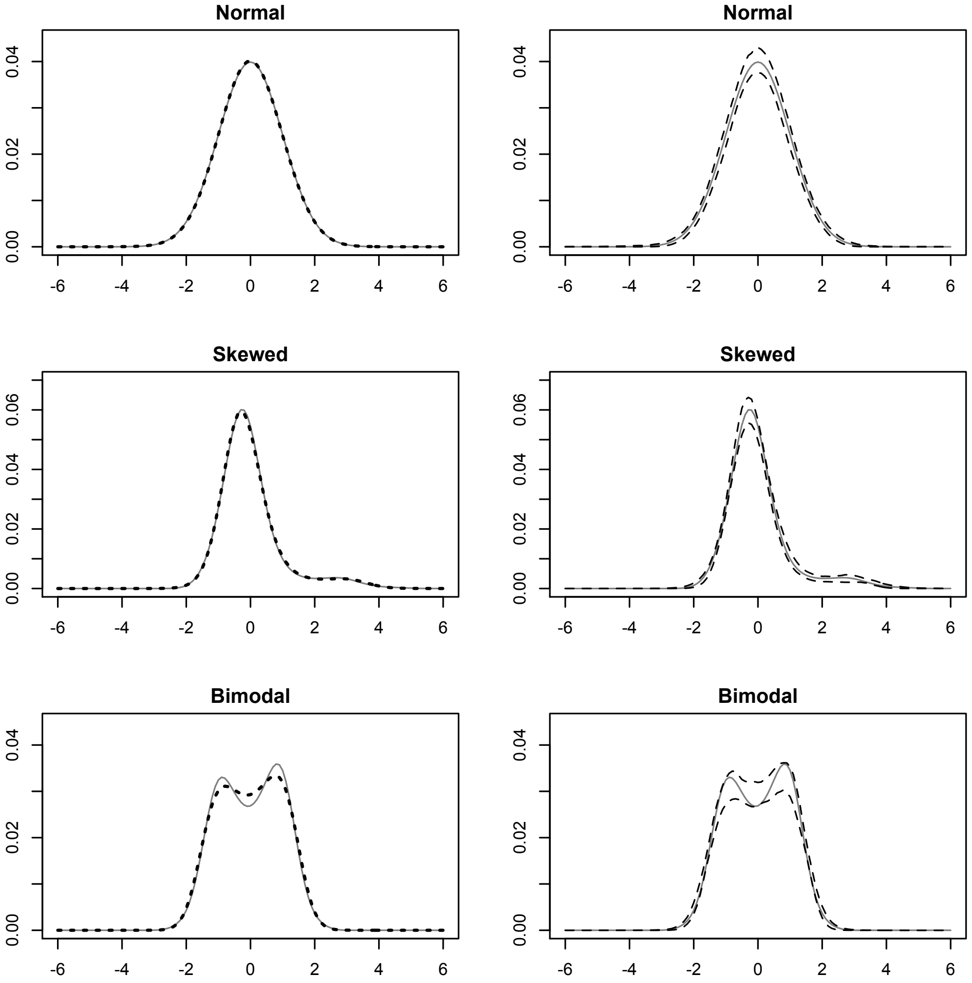
the lowest among the three distributions. While Table 8.2 presents the median ISE values, Figure 8.4 compares the MH-RM and BA-EM ISE values for each distribution by replication. The proximity of the points to the 45° reference line implies that the two methods are yielding comparable density estimates.

Given all of the evidence above, it is clear that the implementations of MH-RM and BA-EM produce very similar point estimates for RC-IRT. Additionally, for MH-RM, figure 8.5 shows the average approximated RC for each condition (left column) and the 95% confidence intervals (right column). For the normal and skewed distributions, the differences between true and average approximated RC are nearly indistinguishable. Also, the 95% confidence intervals for the normal and skewed distributions clearly capture the true curves. For the bimodal distribution, the approximated RCs are less accurate. While the approximated curves are largely bimodal distributions, they fail to capture the full extent of the local extrema.

8.4 Results: Standard Errors from MH-RM

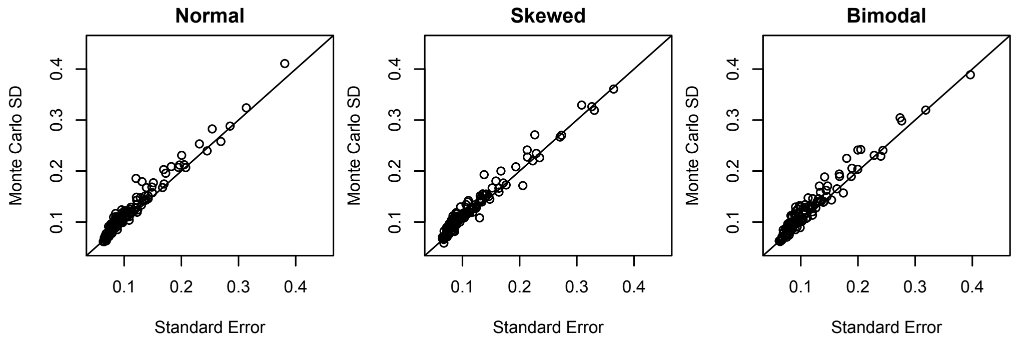
Table 8.3 presents the average standard error and Monte Carlo standard deviation (in parentheses) for all slope parameters across the three distributions. As can be seen, overall, the values are very close to one another. Figure 8.6 presents

Figure 8.5: True densities (gray solid lines) used for the simulations, average RC-IRT estimates (dotted lines), and 95% confidence intervals (dashed lines)



Note. RC-IRT = Ramsay-curve item response theory. Left column: True curves (gray solid lines) and average RC-IRT estimate (dotted line). Right column: True curves (gray solid lines) and 95% confidence interval (dashed lines) for RC-IRT estimate.

Figure 8.6: Average Standard Errors and Monte Carlo Standard Deviations for Item Parameters



the same statistics, but for all item parameters. From the plots, it is clear that the Monte Carlo standard deviations tend to be slightly larger than the average standard errors.

Table 8.3: Average Standard Errors and Monte Carlo Standard Deviations for Slope Parameters

Item	Normal	Skewed	Bimodal
1	0.09 (0.10)	0.10 (0.11)	0.09 (0.10)
2	0.08 (0.08)	0.09 (0.12)	0.08 (0.08)
3	0.09 (0.09)	0.12 (0.13)	0.09 (0.10)
4	0.10 (0.12)	0.12 (0.13)	0.10 (0.09)
5	0.09 (0.10)	0.09 (0.11)	0.09 (0.10)
6	0.09 (0.09)	0.10 (0.11)	0.09 (0.10)
7	0.06 (0.06)	0.07 (0.07)	0.06 (0.06)
8	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)
9	0.07 (0.07)	0.07 (0.07)	0.07 (0.06)
10	0.08 (0.09)	0.08 (0.10)	0.07 (0.08)
11	0.08 (0.09)	0.09 (0.10)	0.08 (0.07)
12	0.08 (0.08)	0.09 (0.09)	0.08 (0.07)
13	0.15 (0.18)	0.17 (0.20)	0.15 (0.14)
14	0.10 (0.10)	0.11 (0.12)	0.09 (0.09)
15	0.11 (0.12)	0.13 (0.14)	0.11 (0.13)
16	0.20 (0.23)	0.21 (0.23)	0.20 (0.20)
17	0.09 (0.09)	0.09 (0.12)	0.09 (0.09)
18	0.08 (0.08)	0.08 (0.11)	0.08 (0.07)
19	0.07 (0.07)	0.07 (0.07)	0.06 (0.06)
20	0.10 (0.12)	0.10 (0.12)	0.10 (0.10)
21	0.11 (0.13)	0.11 (0.14)	0.11 (0.11)
22	0.11 (0.12)	0.11 (0.13)	0.10 (0.10)
23	0.09 (0.10)	0.10 (0.13)	0.09 (0.08)
24	0.09 (0.09)	0.09 (0.11)	0.09 (0.10)
25	0.10 (0.11)	0.10 (0.12)	0.10 (0.11)

Note. Entries are the Monte Carlo averages of estimated standard errors and the Monte Carlo standard deviations (in parentheses) of the estimated parameters.

CHAPTER 9

Empirical Data Analysis

Data used to illustrate RC-IRT via MH-RM with standard errors comes from the Drug Abuse Treatment Outcome Studies (DATOS). DATOS is a national evaluation of treatment effectiveness funded by the National Institues of Drug Abuse. The available sample was quite large, and $N = 2,500$ respondents were randomly selected for the analysis. 11 Likert-type items measuring mental health and emotional distress were analyzed using RC-IRT. As an example, one item asked, "How troubled or distressed (bothered) are you now by emotional or psychological problems?" Respondents could answer with one of the following: *not at all* (0), *somewhat* (1), or *very troubled* (2). Accordingly, higher latent trait scores correspond with greater emotional distress.

To verify that a unidimensional analysis was appropriate, and to provide a comparison for RC-IRT, a standard IRT analysis was carried out using Bock and Aitkin's (1981) EM algorithm in flexMIRT$^{\text{TM}}$ (Cai, 2012). All items were fitted to the graded response model. The resulting RMSEA value was 0.04, suggesting that a unidimensional model fit the data reasonably well. Further results from the standard IRT analysis will be discussed below.

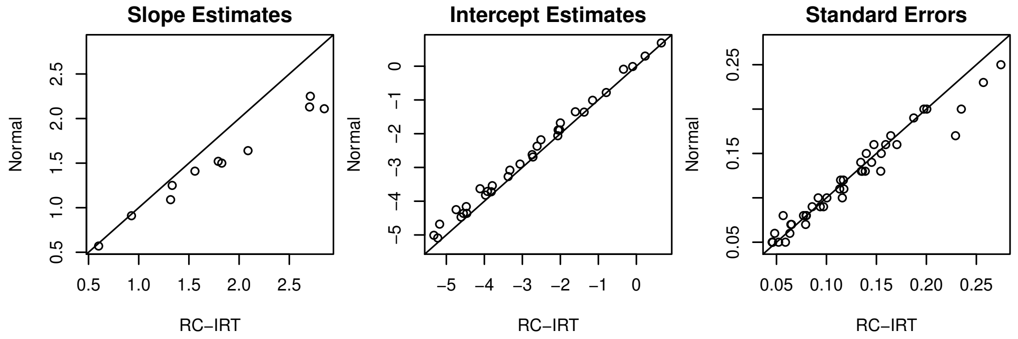
RC-IRT was carried out with maximums of *degree* = 5 and *number of knots* = 6 (see Section 3.1). Model selection was based on the Hannan-Quinn information criteria (HQIC), as recommended by Woods (2007, 2008). The MH-RM specifications were identical to those used for the simulation study (see Section 8).

Table 9.1: Model Comparison Criteria for 11 DATOS Items ($N = 2,500$)

Model	Parameters	-2LogL	AIC	BIC	HQIC
1-3 RC	46	8685.07	8777.07	9044.97	8874.33
Normal	43	8750.19	8836.19	9086.63	8927.11

Note. 1-3 RC = RC-IRT Model with *degree* = 1 and *knots* = 3; Normal = IRT estimation assuming a normal distribution for $g(\theta)$; AIC = Akaike information criteria; BIC= Bayesian information criteria; HQIC = Hannan-Quinn information criteria. All values are less 30,000 to facilitate comparison.

Figure 9.1: Estimates of Item Parameter Standard Errors for 11 DATOS Items



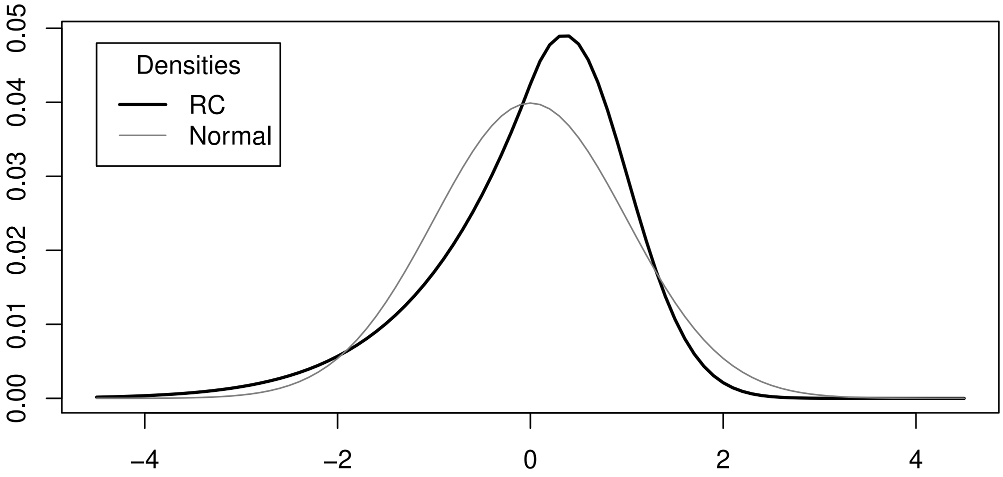
Note. RC-IRT = Ramsay curve IRT; Normal = Standard IRT with Assumption of Normal Density.

9.1 Empirical Analysis Results

For RC-IRT, the 1-3 model (i.e., *degree* = 1 and *knots* = 3) yielded the lowest HQIC, and was thus selected. Table 9.1 displays different comparison criteria for both the RC-IRT and standard IRT models. For all criteria, the 1-3 RC-IRT model has lower values, indicating better fit. Point estimates and standard error estimates for both RC-IRT and standard IRT models are displayed in Figure 9.1. Notably, some standard errors obtained for the 1-3 RC-IRT model are appreciably larger than the corresponding values obtained via traditional analysis.

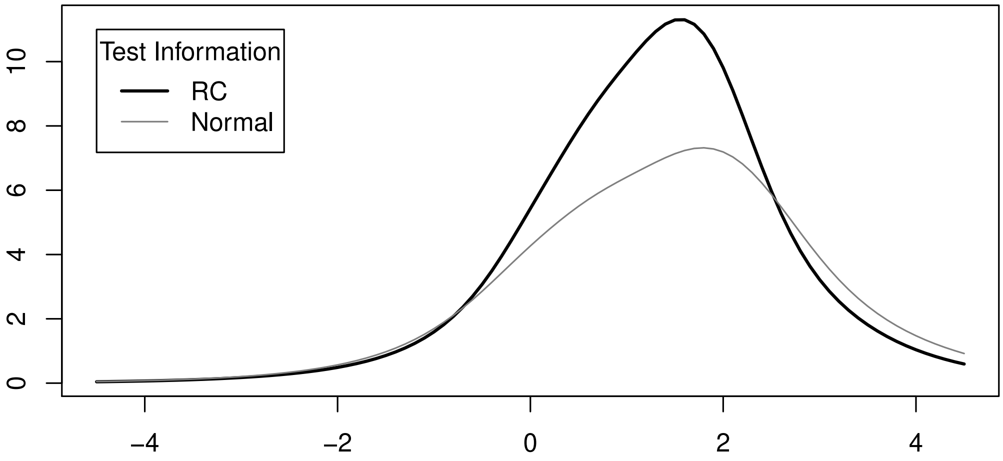
Figure 9.2 shows the estimated RC associated with this scale. Interestingly, the

Figure 9.2: Estimated RC-IRT Curve for 11 DATOS Items ($N = 2,500$)



distribution is skewed left. In comparison to the normal density, the estimated RC indicates a greater number of the respondents are characterized by an exceptional lack of emotional distress. The difference in the two sets of results also manifests itself in the test information curves, shown in Figure 9.3. Of note, neither curve dominates. This implies that the conditional standard error of measurement is not uniformly higher for either analysis. That being said, the most obvious difference in the curves occurs at moderately positive values for the latent trait, where the RC-IRT analysis reveals more information.

Figure 9.3: Estimated Test Information Curve for 11 DATOS Items ($N = 2,500$)



CHAPTER 10

Discussion and Conclusion

The empirical example provides a nice context for discussing some of the advantages of RC-IRT, as well as remaining challenges. First, conventional model selection criteria indices (e.g., BIC, HQIC) ensures that the Ramsay curve may be parsimoniously modelled. In the empirical example, the Ramsay curve was a function of just three parameters. Nevertheless, its shape was clearly nonnormal. Furthermore, the RC-IRT model fit better than the corresponding standard IRT model, as measured by conventional criteria. And, if the criteria reveal that the data do not support non-normality, we can always use the standard model. Thus, from the standpoint of fit, there is little to be lost by using RC-IRT.

On the other hand, by not using RC-IRT, the potential misspecification may lead to other undesirable results. As shown in Figure 9.3, the test information curves from the RC-IRT and standard analyses are clearly different. This discrepancy can have a practical impact on test assembly and item selection, where obtaining a certain test information or conditional standard error of measurment (SEM) curve may be the ultimate goal. Another practical implication involves computer adaptive testing (CAT), where stopping criteria may be based on the conditional SEM. To the extent that RC-IRT models result in smaller conditional SEM values, CAT efficiency may be improved.

Again, in real-world applications, the form of the latent trait distribution is unknown. Thus, treating it as such and estimating its shape from the data is rather compelling. This theoretical argument, along with the practical advantages mentioned above, make a strong case for RC-IRT. Nevertheless, researchers need and prefer methods that do not limit their lines of inquiry. In IRT, the unavailability of standard errors is clearly one such limitation. This research has provided the means to remedy this situation.

Still, other limitations exist. A notable example is that a multidimensional generalization of RC-IRT has not yet been developed. When such a development does occur, MH-RM will be a logical choice for estimation. Unlike BA-EM, MH-RM does not impose artificial ceilings on the dimensionality of a model. To the contrary, the method is, in a sense, designed to address “the curse of dimensionality.” While other methods for implementing standard errors for unidimensional RC-IRT exist, choosing MH-RM lends itself to future generalizations. Hopefully, when this occurs, the current research can serve as a template.

APPENDIX A

Complete Data Log-Likelihood and Derivatives for the Ramsay Curve

The log-likelihood for the Ramsay Curve is

$$l = \sum_{i=1}^{N} \log g(\theta_i | \boldsymbol{\eta}). \tag{A.1}$$

To avoid clutter, let $W(\theta_i) = \mathbf{B}^*(\theta_i)\boldsymbol{\eta}$. Then,

$$\frac{\partial}{\partial \boldsymbol{\eta}} W(\theta_i) = \mathbf{B}^*(\theta_i) \tag{A.2}$$

The first derivatives are

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\eta}} &= \frac{\partial}{\partial \boldsymbol{\eta}} \sum_{i=1}^{N} \log \frac{\exp\left[W(\theta_i)\right]}{C} \\
&= \sum_{i=1}^{N} \left[\frac{\partial}{\partial \boldsymbol{\eta}} \log \exp\left[W(\theta_i)\right] - \frac{\partial}{\partial \boldsymbol{\eta}} \log C \right] \\
&= \sum_{i=1}^{N} \left[\mathbf{B}^*(\theta_i) - \frac{1}{C} \frac{\partial}{\partial \boldsymbol{\eta}} C \right] \\
&= \sum_{i=1}^{N} \left[\mathbf{B}^*(\theta_i) - \frac{1}{C} \sum_{q=1}^{Q} \left(\frac{\partial}{\partial \boldsymbol{\eta}} \exp\left[W(x_q)\right] \right) \right] \\
&= \sum_{i=1}^{N} \left[\mathbf{B}^*(\theta_i) - \sum_{q=1}^{Q} \left(\frac{\exp\left[W(x_q)\right]}{C} \mathbf{B}^*(x_q) \right) \right] \\
&= \sum_{i=1}^{N} \left[\mathbf{B}^*(\theta_i) - \sum_{q=1}^{Q} g(x_q|\boldsymbol{\eta}) \mathbf{B}^*(x_q) \right]
\end{aligned}$$

which corresponds to Equation (14) in Woods and Thissen (2006).

The second derivatives involve $\frac{\partial}{\partial \boldsymbol{\eta}} g(x_q|\boldsymbol{\eta})$, which is:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}} g(x_q|\boldsymbol{\eta}) &= \frac{\partial}{\partial \boldsymbol{\eta}} \frac{\exp\left[W(x_q)\right]}{C} \\
&= \frac{C \left(\frac{\partial}{\partial \boldsymbol{\eta}} \exp\left[W(x_q)\right] \right) - \left(\frac{\partial}{\partial \boldsymbol{\eta}} C \right) \exp\left[W(x_q)\right]}{C^2} \\
&= \frac{C(\exp\left[W(x_q)\right]) \mathbf{B}^*(x_q) - \left(\sum_{q=1}^{Q} \exp\left[W(x_q)\right] \mathbf{B}^*(x_q) \right) \exp\left[W(x_q)\right]}{C^2}
\end{aligned}$$

which, should be noted, does not involve any subscript i. That is, the term is constant across persons.

Finally, the second derivatives are:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'} &= \frac{\partial}{\partial \boldsymbol{\eta}'} \sum_{i=1}^{N} \left[\mathbf{B}^*(\theta_i) - \sum_{q=1}^{Q} g(x_q|\boldsymbol{\eta}) \mathbf{B}^*(x_q) \right] \\
&= \sum_{i=1}^{N} \left[0 - \sum_{q=1}^{Q} \left(\frac{\partial}{\partial \boldsymbol{\eta}} g(x_q|\boldsymbol{\eta}) \mathbf{B}^*(x_q) \right) \right] \\
&= -N \sum_{q=1}^{Q} \left[\left(\frac{\partial}{\partial \boldsymbol{\eta}} g(x_q|\boldsymbol{\eta}) \right) \mathbf{B}^*(x_q)' + g(x_q|\boldsymbol{\eta}) \left(\frac{\partial}{\partial \boldsymbol{\eta}} \mathbf{B}^*(x_q) \right) \right] \\
&= -N \sum_{q=1}^{Q} \left[\left(\frac{\partial}{\partial \boldsymbol{\eta}} g(x_q|\boldsymbol{\eta}) \right) \mathbf{B}^*(x_q)' \right]
\end{aligned}$$

where $\frac{\partial}{\partial \boldsymbol{\eta}} g(x_q|\boldsymbol{\eta})$ is the $v \times 1$ vector given above.

Bibliography

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Borkar, V. S. (2008). *Stochastic approximation: A dynamical systems viewpoint.* Cambridge: Cambridge University Press.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309-329.

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

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

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.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software Internatonal, Inc.

de Boor, C. (2001). *A practical guide to splines* (Revised ed.). New York: Springer-Verlag.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *39*, 1-38.

Diebolt, J., & Ip, E. H. S. (1996). Stochastic EM: method and application. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (p. 259-273). London: Chapman and Hall.

Duncan, K. A., & MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, *8*(1), 41-66.

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.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation.* Unpublished doctoral dissertation, Department of Psychology, University of North Carolina at Chapel Hill.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: 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*(2), 226-233.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*(6), 1087-1092.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359-381.

Patz, R. A., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Polyak, B. T., & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal of control and optimization*, *30*(4), 838-855.

Ramsay, J. O. (2000). Differential equation models for statistical inference. *Canadian Journal of Statistics*, *28*(2), 225-240.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400-407.

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

von Neumann, J. (1951). Various techniques used in connection with random digits. *National Bureau of Standards Applied Mathematics Series*, *12*, 36-38.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, *11*, 253-270.

Woods, C. M. (2007). Ramsay curve IRT for Likert-type data. *Applied Psychological Measurement*, *31*(3), 195-212.

Woods, C. M. (2008). Ramsay curve item response theory for the three-parameter item response theory model. *Applied Psychological Measurement*, *36*(6), 447-465.

Woods, C. M., & Lin, N. (2008). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*(2), 102-117.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The*

Annals of Statistics, *11*(1), 95-103.