Fitting Full-Information Item Factor Models and an Empirical Investigation of Bridge Sampling

Xiao-Li MENG and Stephen SCHILLING

Based on item response theory, Bock and Aitken introduced a method of item factor analysis, termed full-information item factor (FIIF) analysis by Bartholomew because it uses all distinct item response vectors as data. But a limitation of their fitting algorithm is its reliance on fixed-point Gauss–Hermite quadrature, which can produce appreciable numerical errors, especially in high-dimension problems. The first purpose of this article is to offer more reliable methods by using recent advances in statistical computation. Specifically, we illustrate two ways of implementing Monte Carlo Expectation Maximization (EM) algorithm to fit a FIIF model, using the Gibbs sampler to carry out the computation for the E steps. We also show how to use bridge sampling to simulate the likelihood ratios for monitoring the convergence of a Monte Carlo EM, a strategy that is useful in general. Simulations demonstrate substantial improvement over Bock and Aitken's algorithm in recovering known factor loadings in high dimensions. To test our methods, we also apply them to data from LSAT and from a survey on quality of American life, and compare the results to those from the fixed-point approach. Using the FIIF model as a working example, the second purpose of this article is to provide an empirical investigation of the theoretical development of Meng and Wong on bridge sampling, an efficient method for computing normalizing constants. In contrast to importance sampling, which uses draws from one density, bridge sampling uses draws from two (or more) densities and then introduces intermediate densities to "bridge" them. Our empirical investigation confirms the results of Meng and Wong and echoes the empirical evidences documented in computational physics; that is, bridge sampling can reduce simulation errors by orders of magnitude when compared to importance sampling with the same simulation sizes.

KEY WORDS: Factor analysis; Gibbs sampler; Item response theory; Latent variables; Monte Carlo Expectation Maximization algorithm; Normalizing constants; Probit model.

1. INTRODUCTION

Questionnaires and tests purporting to measure attitudes, constructs, and abilities constitute a large part of the data obtained in social survey, educational, and psychological research. Often a subject's responses to a questionnaire or test items are dichotomous and are hypothesized to be intrinsically determined by a small number of underlying latent factors. Within such a theoretical framework, an effective method for exploring underlying factor structures is crucial. Of the various methods that have been proposed, item factor analysis seems to be most researched and used in related fields (e.g., psychology, educational testing). In these fields, common methods of item factor analysis fall into two categories: those based on phi or tetrachoric correlation matrices (e.g., Cristofferson 1975; Muthen 1978) and those fitting latent probit or logit models (e.g., Bartholomew 1987). The full-information item factor (FIIF) approach of Bock and Aitken (1981) belongs to the latter category. In its probit form, the probability of a correct response for a single item under a FIIF model is a normal-ogive function of the subjects score on the hypothetical latent factors. The model is thus a multidimensional normal-ogive item response model or, alternatively, a multivariate latent probit regression model, with the hypothetical latent factors serving as regressors.

The strategies adopted by Bock and Aitken (1981) for fitting a FIIF model were to treat the latent factors as missing data, and then to apply the Expectation Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), with the E step implemented by numerical integrations via fixed-point Gauss–Hermite quadrature. This turns out to be a problematic approach, especially in high-dimensional problems. First, their data augmentation scheme (i.e., augment to include the latent factors) is not ideal for implementing EM, because the corresponding complete-data model is not from an exponential family; thus the numerical inaccuracy in implementing the E step is more likely to create artificial modes for the objective function to be maximized by the M step, a generally not well-recognized problem of using EM with a nonexponential complete-data model (see Meng and Rubin 1992). Second, their M step still needs numerical iterations, as is typical when using EM with a nonexponential complete-data model. Third, the numerical quadrature method used for the E step can be unreliable, especially for high-dimensional integrations. As a result, their algorithm can produce estimates with numerical errors that are large enough to alter the grouping of test items, as we illustrate in Section 3.

The first purpose of this article is to demonstrate how recent advances in computational tools in statistics can offer more reliable algorithms for fitting the FIIF model in multiple dimensions. (For the one-dimensional problem, see Albert 1992.) Specifically, we discuss two applications of the Monte Carlo EM (MCEM) algorithm (see, e.g., Wei and Tanner 1990) with its E step implemented via the Gibbs

Xiao-Li Meng is Assistant Professor, Department of Statistics, University of Chicago, IL 60637. Stephen Schilling is Assistant Professor, Department of Psychology and Human Development, Vanderbilt University, Nashville, TN 37203. The order of the authorship is alphabetical. The majority of this work was done when Schilling was a Ph.D candidate in the Department of Psychology, University of Chicago. This author thanks his advisor R. D. Bock for his guidance. Meng's work was supported in part by National Science Foundation (NSF) Grants DMS 92-04504 and DMS 95-05043. Both authors thank Myles Hollander, an associate editor, and two referees for instructive comments. They also thank A. Kong, M. L. Stein, D. B. Rubin, and W. Wong for helpful conversations. The manuscript was prepared using computer facilities supported in part by several NSF grants awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

sampler (see, e.g., Geman and Geman 1984) to compute the maximum likelihood estimators (MLE's) for the FIIF model; the second application avoids all three problems discussed in the previous paragraph. Moreover, we demonstrate how the method of bridge sampling (see, e.g., Meng and Wong 1996) can be used to determine the convergence of MCEM, a problem encountered in other applications, such as fitting genetic models with MCEM (see, e.g., Guo and Thompson 1992 and Irwin 1994).

EM-type algorithms, including MCEM, and the Gibbs sampler (or, more generally, the iterative simulation) have been the focus of much attention in recent literature (see, e.g., Gelfand and Smith 1990; Little and Rubin 1987; Liu and Rubin 1994; Liu, Wong, and Kong 1994a,b; Meng 1994a; Meng and Pedlow 1992; Meng and Rubin 1991, 1992, 1993, 1994; Tanner 1991; Tanner and Wong 1987; Weeks and Lange 1989; Wei and Tanner 1990; the discussion papers in *Statistical Science* (Nov. 1992) and in *Journal of the Royal Statistical Society,* Ser. B (No. 1, 1993), the papers in the special theme topic on "EM and Related Algorithms" in *Statistica Sinica* (No. 1, 1995), and many other references cited in these papers). Therefore, in this article we list only implementational steps that are directly related to our problem and refer interested readers to the aforementioned literature for general descriptions and discussions of these methods. However, the technique of bridge sampling, which can be viewed as a general formulation of the acceptance ratio method in physics for computing free-energy differences (see, e.g., Bennett 1976 and Voter 1985), is relatively new to statistical researchers, and its general properties have essentially been investigated only theoretically (see, e.g., Gelman and Meng 1994 and Meng and Wong 1996). Therefore, the second purpose of this article is to investigate the performance of bridge sampling in the context of computing likelihood ratios from a FIIF model. Our empirical investigation follows the theoretical development provided by Meng and Wong (1996) and supports their results and predictions, especially on the potential of bridge sampling for substantial reduction of Monte Carlo errors (e.g., by a factor of 5 to 30 in the examples of this article) when compared to importance sampling with the same simulation sizes.

A few disclaimers are in order before we proceed. Recent advances in statistical computation make it easier to fit complicated models, but could also encourage abusive analyses—fitting a complicated model simply because one has the computational ability to do so. We must all make good effort to avoid and discourage such analyses, but that does not imply that we should "hide" powerful computational tools from general practitioners and let them use inferior techniques to fit "wrong" models. Introducing numerical errors into problematic modeling cannot produce more meaningful results but can only make model evaluation and criticism difficult or even impossible. That is why we feel the need to reduce the numerical errors in fitting FIIF models, especially in high dimensions. Like any statistical method, factor analysis can be useful and can be abused (see Maxwell 1983 and citations therein). Our purposes here are purely computational; besides showing how to fit a FIIF model more accurately, we also use this model, as it is simple enough to describe but complicated enough to require these strategies, to illustrate some computational methods that are useful in general (e.g., monitoring convergence of MCEM).

We also do not claim that the fitting algorithms presented here are the best possible ones. Such a claim usually is not very meaningful, because the effectiveness of an algorithm as a tool depends on who is using it. In the hands of a numerically sophisticated analyst, the Gauss–Hermite quadrature method can be made very accurate; in fact the whole EM formulation is not necessarily needed if the analyst has a sophisticated optimization program to directly maximize the intended likelihood function. The methods that we present here, especially the second MCEM, are easier for many users who are not numerically sophisticated and whose goal is not computation.

Finally, although we discuss only point estimates to keep our presentation within a suitable length and to make direct comparisons with work of Bock and Aitken (1981), we do not imply that point estimates are the end of our analysis. In fact, the methods here (e.g., the Gibbs sampler) can be easily modified to obtain the posterior distributions of the parameters of interests (as in Albert and Chib 1993). Moreover, one can use these simulation methods to perform model diagnoses using the posterior predictive check, as discussed by Rubin (1984) and implemented by Gelman, Meng, and Stern (1996) (also see Meng 1994b). We generally feel that using iterative simulation for just computing point estimators is a bit wasteful. We do recognize, however, that there is some reluctance or even resistance to using full Bayesian methods, especially in fields outside of statistics. We feel that it is healthy for the evolution of our ability to deal with real-life problems to be receptive to different perspectives, and thus we have no objection if anyone (including ourselves) is willing to use such "overkill" methods, as long as they indeed solve a statistically relevant computational problem.

2. FITTING A FULL-INFORMATION ITEM FACTOR MODEL

2.1 Model Description

In a FIIF model, subjects' latent factors are tied to the probabilities of correct response for items through the normal-ogive functions. The model can be derived by assuming that Thurstone's d-factor model (Thurstone 1947)

$$y_{ij} = \alpha_{1j} z_{i1} + \cdots + \alpha_{dj} z_{id} + \varepsilon_{ij} \equiv \mathbf{z}_i^\top \alpha_j + \varepsilon_{ij} \qquad (1)$$

describes an unobservable "response process," y_{ij}, instead of an empirically manifest variable. The process yields a correct response for person i to item j when y_{ij} equals or exceeds a threshold parameter γ_j. Assuming that ε_{ij} follow a $N(0, \sigma_j^2)$ distribution, the probability of an item score, $u_{ij} = 1$, indicating a correct response from person i with (unobserved) factor $\mathbf{z}_i = (z_{i1}, \ldots, z_{id})^\top$ is given by

$$\Pr(u_{ij} = 1 | \mathbf{z}_i, \alpha_j, \sigma_j) = \Pr\{y_{ij} \geq \gamma_j | \mathbf{z}_i, \alpha_j, \sigma_j\} = \Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j), \qquad (2)$$

where Φ is the cdf of $N(0,1)$, $\mathbf{a}_j = (a_{1j},\ldots,a_{dj})^\top$ with $a_{mj} = \alpha_{mj}/\sigma_j$ $(m = 1,\ldots,d)$ and $b_j = -\gamma_j/\sigma_j$. Here b_j is called the *item intercept* for item j, γ_j is called the *item difficulty*, a_{mj} is called the *item slope* for factor m, and α_{mj} is called the *item factor loading* for factor m. Notationally, we let $\mathbf{b} = (b_1,\ldots,b_J)^\top$, where $J(>d)$ is the number of items, and let $\mathbf{A}$ be a $d \times J$ matrix whose jth column is $\mathbf{a}_j, j = 1,\ldots,J$. We then call $\theta \equiv \{\mathbf{A},\mathbf{b}\}$ the set of item parameters. Thus a d-factor FIIF model with J items has $J(d+1)$ parameters, and only these parameters are estimable from the observed score matrix $\mathbf{U} = (u_{ij})$; for example, $b_j = -\gamma_j/\sigma_j$ is estimable but σ_j is not.

Given the link function in (2), the FIIF model also assumes that, conditional on a person's (vector) factor $\mathbf{z}$ and model parameter θ, responses to different items are independent of each other. Consequently, given $\mathbf{z}_i$ and θ, the probability of a particular response pattern $\mathbf{u}_i = (u_{i1},\ldots,u_{iJ})^\top$ from the ith person is

$$\Pr(\mathbf{u}_i|\mathbf{z}_i,\theta) = \prod_{j=1}^{J} [\Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)]^{u_{ij}}[1-\Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)]^{1-u_{ij}}.$$

Assuming independence between persons, the likelihood function of θ given responses *and* latent factors from all n persons then is

$$L(\theta|\mathbf{Z},\mathbf{U}) = \prod_{i=1}^{n}\prod_{j=1}^{J} [\Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)]^{u_{ij}} \times [1-\Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)]^{1-u_{ij}}, \quad (3)$$

where $\mathbf{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_n)^\top$ is the $n \times d$ (latent) factor matrix.

Because $\mathbf{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_n)^\top$ is not observable, the FIIF model assumes $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are independently and identically distributed according to the standard d-variate normal distribution, $N_d(0,\mathbf{I})$. Integrating over $\mathbf{z}_i$'s in (3) yields the actual likelihood from the FIIF model given $\mathbf{U}$,

$$L(\theta|\mathbf{U}) = \prod_{i=1}^{n} E_{\mathbf{z}}\left\{\prod_{j=1}^{J} [\Phi(\mathbf{z}^\top \mathbf{a}_j + b_j)]^{u_{ij}} \times [1-\Phi(\mathbf{z}^\top \mathbf{a}_j + b_j)]^{1-u_{ij}}\right\}, \quad (4)$$

where the expectation, $E_{\mathbf{z}}$, is with respect to $\mathbf{z} \sim N_d(0,\mathbf{I})$. From (4), we see that subjects with the same response pattern, $\mathbf{u}_i$, contribute equally to the likelihood, and thus (4) can be simplified to

$$L(\theta|\mathbf{U}) = \prod_{i=1}^{n_0}\left\{E_{\mathbf{z}}\left[\prod_{j=1}^{J}[\Phi(\mathbf{z}^\top \mathbf{a}_j + b_j)]^{u_{ij}} \times [1-\Phi(\mathbf{z}^\top \mathbf{a}_j + b_j)]^{1-u_{ij}}\right]\right\}^{s_i}, \quad (5)$$

where, for notational simplicity, i now indexes distinct response pattern, s_i is the number of the subjects who share $\mathbf{u}_i$, and $n_0 (\le \min\{n, 2^J\})$ denotes the number of the distinct response patterns. The importance of reexpressing (4) as (5) is to recognize that the FIIF model cannot distinguish subjects with the same response pattern, allowing us to assign one latent variable $\mathbf{z}$ to all subjects that share a response pattern and thus increase computational efficiency. Therefore, $\mathbf{U}$ is subsequently (notationally) reduced to a $n_0 \times J$ score matrix corresponding to the distinct response patterns, and $\mathbf{Z}$ is reduced accordingly. Maximizing $L(\theta|\mathbf{U})$ in (5)—that is, finding the MLE of θ for a given $\mathbf{U}$—is the first objective of this article.

2.2 Two EM Implementations for Estimating the Item Parameters

Bock and Aitken (1981) originally proposed estimating θ via the EM algorithm by treating $\{\mathbf{U},\mathbf{Z}\}$ as the "complete" data and the latent variable $\mathbf{Z}$ as the "missing" data. Treating latent variables as missing data is a common strategy in the EM literature (see, e.g., Meng and Rubin 1996 and Rubin and Thayer 1982), and the usefulness of such a strategy lies in whether the ML estimation given the complete data would be easier. In the current setting, if both $\mathbf{U}$ and $\mathbf{Z}$ were observed, then maximizing the complete-data likelihood, given by (3), is equivalent to maximizing J *separate* functions, each involving only $d+1$ parameters, namely $\{\mathbf{a}_j, b_j\}, j = 1,\ldots,J$. In contrast, directly maximizing the $L(\theta|\mathbf{U})$ of (5) requires handling all $J(d+1)$ parameters simultaneously. In typical FIIF applications, J can range from 10 to 1,000, yielding up to thousands of item parameters even with only a few factors (e.g., 4). Implementing standard numerical procedures, such as the Newton–Raphson algorithm, for very high-dimensional problems is computationally intractable. Thus Bock and Aitken's EM formulation simplifies the task for maximization.

Unfortunately, in this setting the simplicity of the M step comes partially at the expense of the computation of the E step. The E step requires taking the conditional expectation of $L(\theta|\mathbf{Z},\mathbf{U})$ over $\mathbf{Z}$ given $\mathbf{U}$ and the parameter estimate from the previous M step, $\theta^{(t)}$, where t indexes iteration. Using the notation of Dempster et al. (1977), the E step must compute

$$Q(\theta|\theta^{(t)}) = \sum_{j=1}^{J}\left\{\sum_{i=1}^{n_0} s_i\{u_{ij}E[\log\Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)|\mathbf{u}_i,\theta^{(t)}] + (1-u_{ij})E[\log(1-\Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j))|\mathbf{u}_i,\theta^{(t)}]\}\right\}, \quad (6)$$

where the expectations are with respect to $f(\mathbf{z}_i|\mathbf{u}_i,\theta^{(t)}), i = 1,\ldots,n_0$. Bock and Aitken (1981) proposed using fixed-point Gauss–Hermite quadrature to compute these d-dimensional expectations (integrations). But even with moderate d (e.g., 5) this approach becomes unwieldy, as the

number of quadrature points required increases exponentially with the number of factors. Furthermore, as is well known, the accuracy of numerical integrations diminishes rapidly with the dimension (see Sec. 3). On the other hand, Monte Carlo simulations are relatively stable in high dimensions, which leads to the idea of implementing the E step via Monte Carlo.

Before we discuss implementing the E step via simulation, we point out that there is a different—in fact, more ideal—way of implementing EM for this problem. The idea here is to further augment $\{\mathbf{U}, \mathbf{Z}\}$ to $\{\mathbf{U}, \mathbf{Z}, \mathbf{X}\}$, where the new "data" $\mathbf{X} = (x_{ij})$ are assumed to have independent (with respect to both i and j) conditional normal distributions given $\mathbf{Z}$:

$$x_{ij}|\mathbf{z}_i, \mathbf{a}_j, b_j, \sim N(\mathbf{z}_i^\top \mathbf{a}_j + b_j, 1).$$

This data augmentation scheme is inspired by (2), where we have

$$\Pr(u_{ij} = 1|\mathbf{z}_i, \mathbf{a}_j, b_j) = \Pr\{x_{ij} \geq 0|\mathbf{z}_i, \mathbf{a}_j, b_j\} = \Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j).$$

In other words, we define x_{ij} such that $u_{ij} = 1_{(x_{ij} \geq 0)}$, where 1_A is the indicator function of a set A. Such a data-augmentation scheme has been used for handling similar probit models (see, e.g., Albert and Chib 1993 and McCulloch and Rossi 1991). Implementing the M step in this setting is almost trivial. If both $\mathbf{X}$ and $\mathbf{Z}$ were observed, then the MLE's of $\mathbf{A}$ and $\mathbf{b}$ can be obtained via J separate linear regressions, because

$$\tilde{\mathbf{x}}_j|\mathbf{a}_j, b_j, \mathbf{Z} \sim N_n(\mathbf{Z}\mathbf{a}_j + b_j\mathbf{1}, \mathbf{I}), \quad \text{independently for} \quad j = 1, \ldots, J,$$

where $\tilde{\mathbf{x}}_j = (x_{1j}, \ldots, x_{n_0 j})^\top$ and $\mathbf{1}$ is a vector of 1s. Furthermore, here the E step involves computing only the conditional expectations of the complete-data sufficient statistics, $\{\mathbf{Z}^\top\mathbf{Z}, \mathbf{Z}^\top\mathbf{X}, \mathbf{1}^\top\mathbf{X}, \mathbf{1}^\top\mathbf{Z}\}$, given $\mathbf{U}$ and $\theta^{(t)}$. Consequently, the current EM is less vulnerable to numerical instability.

To facilitate the computation, we let $\mathbf{x}_i^\top$ be the ith row of $\mathbf{X}, i = 1, \ldots, n_0$, and $\mathbf{V} = (\mathbf{I} + \mathbf{A}\mathbf{A}^\top)^{-1}$. Then, under our model specification for $\{\mathbf{X}, \mathbf{Z}\}$, we have

$$\mathbf{x}_i|\theta \sim \text{iid } N_J(\mathbf{b}, \mathbf{I} + \mathbf{A}^\top\mathbf{A}) \tag{7}$$

and

$$\mathbf{z}_i|\mathbf{x}_i, \theta \sim \text{ind. } N_d(\mathbf{V}\mathbf{A}(\mathbf{x}_i - \mathbf{b}), \mathbf{V}), \qquad i = 1, \ldots, n_0. \tag{8}$$

We further let $\mathbf{e}(\mathbf{u}_i|\theta) = E[(\mathbf{x}_i - \mathbf{b})|\mathbf{u}_i, \mathbf{b}, \mathbf{A}]$ and $\mathbf{D}(\mathbf{u}_i|\theta) = E[(\mathbf{x}_i - \mathbf{b})(\mathbf{x}_i - \mathbf{b})^\top|\mathbf{u}_i, \mathbf{b}, \mathbf{A}]$. Then it is easy to verify from (7) and (8) (recall that s_i is the number of subjects who share $\mathbf{u}_i$) that

$$E[\mathbf{1}^\top\mathbf{X}|\mathbf{U}, \theta] = \sum_{i=1}^{n_0} s_i[\mathbf{e}^\top(\mathbf{u}_i|\theta) + \mathbf{b}^\top], \tag{9}$$

$$E[\mathbf{1}^\top\mathbf{Z}|\mathbf{U}, \theta] = \left[\sum_{i=1}^{n_0} s_i\mathbf{e}^\top(\mathbf{u}_i|\theta)\right]\mathbf{A}^\top\mathbf{V}^\top, \tag{10}$$

$$E[\mathbf{Z}^\top\mathbf{X}|\mathbf{U}, \theta] = \mathbf{V}\mathbf{A}\sum_{i=1}^{n_0} s_i[\mathbf{D}(\mathbf{u}_i|\theta) + \mathbf{e}(\mathbf{u}_i|\theta)\mathbf{b}^\top], \tag{11}$$

and

$$E[\mathbf{Z}^\top\mathbf{Z}|\mathbf{U}, \theta] = \mathbf{V} + \mathbf{V}\mathbf{A}\left[\sum_{i=1}^{n_0} s_i\mathbf{D}(\mathbf{u}_i|\theta)\right]\mathbf{A}^\top\mathbf{V}^\top. \tag{12}$$

Computing $\mathbf{e}(\mathbf{u}|\theta)$ and $\mathbf{D}(\mathbf{u}|\theta)$ is equivalent to computing the first two moments of a truncated multivariate normal with general covariance structures, because $u_{ij} = 1_{(x_{ij} \geq 0)}$. This is known to be a difficult task, especially when the dimension is high, and has been often carried out via Monte Carlo simulations (see, e.g., Stein 1992); the Gibbs sampler described in Section 2.3 is an easy approach to conducting such a simulation.

For general users, we recommend the second EM, because its M step is trivial to implement and its E step is also simpler and more stable. It is particularly attractive if a user has access to subroutines that compute (accurately) the moments of a truncated multivariate normal as functions of the truncating points; the recurrence relations given by Gupta and Tracy (1976) can be useful for establishing such subroutines. The disadvantage of the second EM is that it converges slower, because it has a higher fraction of missing information (Dempster et al. 1977; Meng and Rubin 1991). For the examples in Section 3, we implemented both EM algorithms (using the Gibbs sampler of Sec. 2.3 to perform the E steps) and found that the second EM takes about twice as many iterations to converge as the first EM. But compared to the effort and care one must take to implement the first EM, the trade-off is well worthwhile. However, to compare fairly with Bock and Aitken's (1981) approach, all of the numerical results presented in Section 3 are based on the first EM; the results from the second EM are the same to the extent allowed by the Monte Carlo errors introduced at the E steps.

2.3 Implementing the *E* Step via the Gibbs Sampler

To implement a Monte Carlo E step, which simulates the expected complete-data log-likelihood function (e.g., (6)), we need draws from $f(\mathbf{Z}|\mathbf{U}, \theta)$ for the first EM and from $f(\mathbf{X}, \mathbf{Z}|\mathbf{U}, \theta)$ (or only from $f(\mathbf{X}|\mathbf{U}, \theta)$) for the second EM. Directly making draws from these conditional distributions is difficult, but making draws from $f(\mathbf{Z}|\mathbf{X}, \mathbf{U}, \theta)$ is trivial, because $f(\mathbf{Z}|\mathbf{X}, \mathbf{U}, \theta) = f(\mathbf{Z}|\mathbf{X}, \theta)$, which is given by (8). On the other hand, drawing from $f(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \theta)$ is the same as drawing from $n_0 \times J$ univariate truncated normal distributions, because given $\mathbf{Z}, \mathbf{U}$, and θ, x_{ij}'s are independent $N(\mathbf{z}_i^\top\mathbf{a}_j + b_j, 1)$ truncated at the left by zero if $u_{ij} = 1$ and at the right by zero if $u_{ij} = 0$. This immediately suggests using the Gibbs sampler to iterate between draws from $f(\mathbf{Z}|\mathbf{X}, \mathbf{U}, \theta)$ and from $f(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \theta)$ until the equilibrium distribution $f(\mathbf{X}, \mathbf{Z}|\mathbf{U}, \theta)$ is reached. In our implementations, this Gibbs sampler mixed very fast, with autocorrelations typically decaying to near zero after three iterations. In the examples in Section 3, we discarded the first ten iterations, and then chose every fifth iteration afterward until the required number of draws, K (e.g., 25), was reached. We

used every fifth draw to produce approximately K independent draws from the desired joint density $f(\mathbf{X}, \mathbf{Z}|\mathbf{U}, \theta)$. It is not necessary to achieve independence to have consistent Gibbs sampler estimates, but we did so because the Gibbs sampler here is cheap compared to the subsequent function evaluations, especially for the first EM.

Once we have K draws, $\{\mathbf{X}_k^{(t)}, \mathbf{Z}_k^{(t)}, k = 1, \ldots, K\}$, from $f(\mathbf{X}, \mathbf{Z}|\mathbf{U}, \theta^{(t)})$, where $\theta^{(t)} = \{\mathbf{A}^{(t)}, \mathbf{b}^{(t)}\}$ is from the tth M step, we use them to form Monte Carlo estimates for the expected complete-data log-likelihood function or, equivalently, the expected complete-data sufficient statistics for the second EM. Specifically, for the first EM, we replace (6) by

$$\hat{Q}_K(\theta|\theta^{(t)}) = \sum_{j=1}^{J}\left\{\sum_{i=1}^{n_0} s_i\left\{\frac{u_{ij}}{K}\sum_{k=1}^{K}\log\Phi([\mathbf{z}_{k,i}^{(t)}]^\top\mathbf{a}_j + b_j) + \frac{1-u_{ij}}{K}\sum_{k=1}^{K}\log[1-\Phi([\mathbf{z}_{k,i}^{(t)}]^\top\mathbf{a}_j + b_j)]\right\}\right\},$$

where $\mathbf{Z}_k^{(t)} = (\mathbf{z}_{k,1}^{(t)}, \ldots, \mathbf{z}_{k,n_0}^{(t)})^\top$. The next M step then maximizes this estimated log-likelihood function to obtain the next iterate $\theta^{(t+1)}$; in our examples we use a Newton–Raphson subroutine to maximize each of the J terms summing up to $\hat{Q}_K(\theta|\theta^{(t)})$.

For the second EM, there are two ways of using the draws, $\{\mathbf{X}_k^{(t)}, \mathbf{Z}_k^{(t)}, k = 1, \ldots, K\}$, for implementing the E step. The first way is simply to replace the complete-data sufficient statistics, $\{\mathbf{Z}^\top\mathbf{Z}, \mathbf{Z}^\top\mathbf{X}, \mathbf{1}^\top\mathbf{X}, \mathbf{1}^\top\mathbf{Z}\}$, by their corresponding sample averages from $\{\mathbf{X}_k^{(t)}, \mathbf{Z}_k^{(t)}, k = 1, \ldots, K\}$. The second way is to use only $\{\mathbf{X}_k^{(t)}, k = 1, \ldots, K\}$ to compute

$$\hat{\mathbf{e}}(\mathbf{u}_i|\theta^{(t)}) = \frac{1}{K}\sum_{k=1}^{K}[\mathbf{x}_{i,k}^{(t)} - \mathbf{b}^{(t)}]$$

and

$$\hat{\mathbf{D}}(\mathbf{u}_i|\theta^{(t)}) = \frac{1}{K}\sum_{k=1}^{K}[\mathbf{x}_{i,k}^{(t)} - \mathbf{b}^{(t)}][\mathbf{x}_{i,k}^{(t)} - \mathbf{b}^{(t)}]^\top,$$

and substitute them for $\mathbf{e}(\mathbf{u}_i|\theta)$ and $\mathbf{D}(\mathbf{u}_i|\theta)$ in (9)–(12), where $\mathbf{b}, \mathbf{A}$, and $\mathbf{V}$ are also replaced by their counterparts from $\theta^{(t)}$. We recommend the second method; whenever possible, one should use "Rao-Blackwellization". The only disadvantage of the second approach is the requirement of inverting a $d \times d$ matrix to produce $\mathbf{V}$. But in typical applications, the number of factors, d, is well within, say, 10 (the point of FIIF is to identify a few key factors), and inverting matrices of such sizes is an easy task now.

We point out that if one has access to a subroutine that can make draws directly from a truncated multivariate normal distribution with arbitrary mean and covariance matrix, then one should replace the foregoing Gibbs sampler scheme by a noniterative simulation scheme (although the subroutine itself may be based on an iterative scheme). Specifically, one can first use such a subroutine to draw from $f(\mathbf{X}|\mathbf{U}, \theta)$, a truncated multivariate normal as implied by (7), and then, given the drawn $\mathbf{X}$, draw $\mathbf{Z}$ from $f(\mathbf{Z}|\mathbf{X}, \mathbf{U}, \theta)$ according to (8). However, drawing directly from a truncated multivariate normal distribution with arbitrary mean and covariance matrix is not as easy as one might first think, especially when the truncated region is small and/or the dimension is high, both of which occur for the FIIF applications. In contrast, when conditioning on both $\mathbf{Z}$ and $\mathbf{U}$, all elements of $\mathbf{X}$ are mutually independent, and the problem then becomes to simulate from many *univariate* truncated normal distributions. In our implementation we used an adaptive rejection method for drawing from an univariate truncated normal. When the truncated region was large, we used straightforward rejection, and when the truncated part was in a small tail, we used an exponential envelope. Thus the Gibbs sampler allows us to transform a high-dimensional problem to many one-dimensional problems. In fact, the Gibbs sampler approach described here provides an efficient way (because it mixes fast) to simulate from a truncated multivariate normal distribution with covariance matrix in the form of $\mathbf{I} + \mathbf{B}^\top\mathbf{B}$; see (7).

2.4 Determining Convergence of MCEM via Bridge Sampling

Because of the simulation variability introduced at its E step, an MCEM sequence typically exhibits random fluctuation around a stationary point θ^*, even on convergence. Thus, unlike standard implementations of EM, it is generally difficult to terminate an MCEM iteration according to the criteria that the (relative) differences between consecutive iterates are within a desired level. Wei and Tanner (1990) recommended plotting $\theta^{(t)}$ against t, terminating the EM iterations when the plot exhibits fluctuation around some line $\theta = \theta^*$. But hundreds or even thousands of item parameters are common in FIIF applications, making plotting individual components of $\theta^{(t)}$ impractical. A more practical graphical approach is to plot some functions of $\theta^{(t)}$ or of both $\theta^{(t)}$ and $\theta^{(t+1)}$. For instance, in our examples we plot $\delta_0^{(t)} = \max_j\{|b_j^{(t+1)} - b_j^{(t)}|\}$ and $\delta_m^{(t)} = \max_j\{|a_{mj}^{(t+1)} - a_{mj}^{(t)}|\}$ against t for $m = 1, \ldots, d$ (see Figs. 2 and 4). Another function is the log-likelihood value or the difference of the consecutive log-likelihood values. For inference purposes, it is always desirable to evaluate the likelihood. Furthermore, the celebrated feature of EM—namely, the monotonic convergence in likelihood—allows us to detect implementational errors and/or numerical inaccuracy if a likelihood plot violates the monotonicity.

For MCEM, monitoring the likelihood values is not a trivial task. First, even without implementation or numerical errors, the log-likelihood can still "zigzag" along the iterates (although it should show an increasing trend) and can fluctuate around an asymptote, to the extent allowed by simulation variability. Second, the fact that we must use simulation to implement the E step generally implies that we cannot evaluate the actual likelihood analytically and thus must numerically compute the log-likelihood values needed for plotting. The second problem is particularly problematic, because if we cannot compute those likelihood values

with good accuracy, then we will not be able to tell from the log-likelihood plot whether any large fluctuation is due to nonconvergence of MCEM, to implementation errors, or to large numerical errors in computing the likelihood values. We previously had faced such a problem when we used the conventional importance sampling approach to estimate the likelihood values, whose large variabilities made it impossible to detect the convergence of an MCEM (e.g., Fig. 5). The bridge sampling approach presented here enabled us to resolve this problem. Because this approach is relatively new and can be applied to monitor convergence of MCEM in general, we give a general description and treat the application to FIIF as a special case.

Let $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\}$ be the complete data under an EM setting, where $\mathbf{Y}_{\text{obs}}$ is the actual observed data and $\mathbf{Y}_{\text{mis}}$ is the missing data (or latent variable). In a typical MCEM setting, the complete-data likelihood function $L(\theta|\mathbf{Y}) = f(\mathbf{Y}|\theta)$ is easy to evaluate, but the observed-data likelihood $L(\theta|\mathbf{Y}_{\text{obs}}) = f(\mathbf{Y}_{\text{obs}}|\theta)$ is not. In addition, we can make draws from the conditional density $f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \theta)$, as required by the Monte Carlo E step. Now the simple identity

$$f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \theta) = \frac{f(\mathbf{Y}|\theta)}{f(\mathbf{Y}_{\text{obs}}|\theta)} = \frac{L(\theta|\mathbf{Y})}{L(\theta|\mathbf{Y}_{\text{obs}})} \tag{13}$$

indicates that computing $L(\theta|\mathbf{Y}_{\text{obs}})$ is equivalent to computing the *normalizing constant* of the conditional density of $f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \theta)$, by viewing $L(\theta|\mathbf{Y})$ as an unnormalized density for $\mathbf{Y}_{\text{mis}}$. Because in monitoring the convergence of a likelihood only the changes in likelihood values are of interest, this is a setting where we want to evaluate ratios of normalizing constants of densities from which we have draws. This setting appears in a variety of problems, as discussed by Meng and Wong (1996).

Common methods for such a problem are to use the following identity (e.g., Geyer and Thompson 1992):

$$\frac{c_2}{c_1} = E_1\left[\frac{q_2(\omega)}{q_1(\omega)}\right], \tag{14}$$

where $p_l(\omega) = q_l(\omega)/c_l, \omega \in \Omega_l, l = 1, 2$ are two densities; $q_l(\omega), l = 1, 2$ are easy to evaluate; and $c_l, l = 1, 2$ are unknown normalizing constants. In (14) the expectation is with respect to the first density, p_1, and with draws $\{\omega_{11}, \ldots, \omega_{1n_1}\}$ from p_1, we can use the sample average of $\{q_2(\omega_{1j})/q_1(\omega_{1j}), j = 1, \ldots, n_1\}$ to estimate the desired ratio, $r = c_2/c_1$. Because this is a special case of importance sampling, the variance of the estimator increases rapidly with the chi-squared distance between p_2 and p_1.

In standard importance sampling applications, we often have draws from only one trial density, and thus the aforementioned problem is inevitable once the trial density is chosen. In the current setting, however, it is often equally easy to make draws from q_1 as from q_2. For the FIIF model, making a draw from $f(\mathbf{X}, \mathbf{Z}|\mathbf{U}, \theta)$ at $\theta = \theta_1$ and at $\theta = \theta_2$ requires the same amount of computation. Motivated by this observation, Meng and Wong (1996) suggested constructing estimators of r based on the identity

$$r = \frac{c_2}{c_1} = \frac{E_1[q_2(\omega)\alpha(\omega)]}{E_2[q_1(\omega)\alpha(\omega)]}, \tag{15}$$

where α is an arbitrary function satisfying $0 < |\int_{\Omega_1 \cap \Omega_2} \alpha(\omega)p_1(\omega)p_2(\omega)\, d\omega| < \infty$. Identity (15) allows us to use draws from both p_1 and p_2 and to use the α function as a link to "bridge" the two densities, thus achieving better precision compared to estimators based on (14); see Section 4 for further discussion.

For reasons discussed later, we chose $\alpha = (\sqrt{q_1 q_2})^{-1}$ for our FIIF applications. To be specific, consider the first EM. We first notice from (5) that an actual likelihood ratio can be expressed as

$$\frac{L(\theta_2|\mathbf{U})}{L(\theta_1|\mathbf{U})} = \prod_{i=1}^{n_0}\left[\frac{L(\theta_2|\mathbf{u}_i)}{L(\theta_1|\mathbf{u}_i)}\right]^{s_i}. \tag{16}$$

Now viewing $f(\mathbf{z}|\mathbf{u}_i, \theta_l)$ as p_l and $L(\theta_l|\mathbf{u}_i)$ as $c_l, l = 1, 2$, we can apply bridge sampling to estimate each individual ratio in (16) by using (see (3))

$$L(\theta_l|\mathbf{z}, \mathbf{u}_i) = \prod_{j=1}^{J}[\Phi(\mathbf{z}^\top \mathbf{a}_j^{(l)} + b_j^{(l)})]^{u_{ij}} \times [1 - \Phi(\mathbf{z}^\top \mathbf{a}_j^{(l)} + b_j^{(l)})]^{1-u_{ij}} \tag{17}$$

as the unnormalized densities, q_l, where the superscripts of $\mathbf{a}_j$ and b_j correspond to the subscript of θ. Now suppose that we have $\tilde{K}$ draws from each of $f(\mathbf{z}|\mathbf{u}_i, \theta_l)$, denoted by $\{\mathbf{z}_k^{(l)}, k = 1, \ldots, \tilde{K}\}$. Then a bridge sampling estimate for the ith ratio, $L(\theta_2|\mathbf{u}_i)/L(\theta_1|\mathbf{u}_i)$, using $\alpha = (\sqrt{q_1 q_2})^{-1}$, is given by

$$\hat{r}_i(\theta_2, \theta_1) = \frac{\sum_{k=1}^{\tilde{K}}\left[\frac{L(\theta_2|\mathbf{z}_k^{(1)}, \mathbf{u}_i)}{L(\theta_1|\mathbf{z}_k^{(1)}, \mathbf{u}_i)}\right]^{1/2}}{\sum_{k=1}^{\tilde{K}}\left[\frac{L(\theta_1|\mathbf{z}_k^{(2)}, \mathbf{u}_i)}{L(\theta_2|\mathbf{z}_k^{(2)}, \mathbf{u}_i)}\right]^{1/2}}, \tag{18}$$

where $L(\theta|\mathbf{z}, \mathbf{u}_i)$ is from (17). The estimate for the log of the ratio of (16) is then

$$\hat{g}_K(\theta_2, \theta_1) = \sum_{i=1}^{n_0} s_i \log \hat{r}_i(\theta_2, \theta_1). \tag{19}$$

In determining the convergence of a MCEM, we plot $\hat{g}_K(\theta^{(t+1)}, \theta^{(t)})$ against t, the index for iteration. If the plot shows a curve converging from above to zero (because EM should increase the likelihood), with a fluctuation that can be expected from the simulation sizes, then this is an indication that an approximate convergence has been achieved (see the examples in Sec. 3). We emphasize that with MCEM, approximate convergence is all we can obtain. Fortunately, this is almost always enough for the purpose of statistical inference.

Table 1. Estimates and True Values for Example 1

	Factor 1			Factor 2		
Item	*GHEM*	*MCEM*	*True*	*GHEM*	*MCEM*	*True*
1	−.543	−.519	−.52	−.274	−.321	−.34
2	−.595	−.567	−.52	−.339	−.388	−.34
3	−.545	−.583	−.52	−.265	−.290	−.34
4	−.520	−.556	−.52	−.318	−.354	−.34
5	−.526	−.559	−.52	−.272	−.307	−.34
6	−.530	−.565	−.52	−.334	−.368	−.34
7	−.526	−.554	−.52	−.344	−.376	−.34
8	−.497	−.530	−.52	−.361	−.391	−.34
9	−.404	−.513	−.52	.409	.388	.34
10	−.397	−.482	−.52	.348	.327	.34
11	−.409	−.510	−.52	.388	.364	.34
12	−.502	−.602	−.52	.406	.359	.34
13	−.386	−.492	−.52	.377	.341	.34
14	−.350	−.451	−.52	.435	.400	.34
15	−.371	−.476	−.52	.334	.321	.34
16	−.385	−.488	−.52	.406	.380	.34
17	.391	.476	.52	−.334	−.299	−.34
18	.455	.557	.52	−.416	−.369	−.34
19	.462	.569	.52	−.361	−.348	−.34
20	.446	.549	.52	−.377	−.350	−.34
21	.445	.554	.52	−.384	−.366	−.34
22	.404	.508	.52	−.373	−.346	−.34
23	.407	.497	.52	−.436	−.419	−.34
24	.425	.530	.52	−.311	−.295	−.34
25	.524	.565	.52	.253	.284	.34
26	.491	.521	.52	.308	.330	.34
27	.498	.533	.52	.298	.326	.34
28	.447	.479	.52	.285	.300	.34
29	.476	.509	.52	.308	.327	.34
30	.475	.504	.52	.373	.394	.34
31	.507	.545	.52	.281	.316	.34
32	.510	.534	.52	.321	.356	.34
ASE	.0068	.0013		.0026	.0013	

3. SIMULATION STUDIES AND REAL-DATA ILLUSTRATIONS

3.1 Example 1: A Two-Factor Simulation

In our first simulation study, 1,000 response patterns for a 32-item test were simulated according to the FIIF specification with $\sigma_j \equiv 1$, all item difficulties zero, and true factor loadings provided in Table 1. These factor loadings represent an orthogonal factor pattern derived from a Hadamard matrix, with one dominating factor. This choice was made because such a regular pattern of item factor loadings and difficulties should be most favorable for Bock and Aitken's Gauss–Hermite EM (GHEM). For the simulated data, GHEM was implemented with 25 quadrature points and MCEM was applied with 25 (Gibbs sampler) samples at each E step; such choices roughly equate the computational load in our experiences, and hereafter we refer to the sizes of the samples from the Gibbs sampler also as "points" to indicate the equivalence. As Table 1 shows, MCEM more accurately reproduces the true factor loadings than GHEM. As a simple overall *numerical* measure of the performance of each method, we take the average over items of the sum of squared residuals from the true factor loadings. The MCEM produces average squared errors (ASE's) 1/5 and 1/2 of those of GHEM for the first factor and the second factor. The errors from the two methods would have been identical had there been no simulation variability (for MCEM) or numerical inaccuracy (for GHEM), because both methods intend to compute the same MLE; converging to two different modes is very unlikely here, as we used the same EM formulation and started with the same initial value—the MINRES factor loadings (e.g., Harman 1976) obtained from the smoothed tetrachoric correlation matrix.

From Table 1, we also notice that the first factor loading from MCEM is always greater (in magnitude) than the second for all items, consistent with the true factor loadings. In contrast, items 14, 16, and 23 show a higher loading on the second factor from GHEM estimates. These reversals of order in factor loadings often have greater practical import than mere inaccuracy, because they can lead to different grouping of the items under study (see, e.g., Example 4). Of course, in real applications one should take into account the inherent variabilities in estimates when comparing the magnitude of factor loadings (but inaccurate calculations of MLE's can also lead to incorrect variance–covariance calculations), but the reversals that we see here are clearly due to the defect of a computation method and thus must be eliminated before one can reach any meaningful conclusion.

It is possible in low dimensions and with a moderate number of items to achieve comparable computational accuracy for GHEM by increasing the number of quadrature points. In this example, increasing the number of Gauss–Hermite quadrature points to 100 resulted in ASE's of 1.5×10^{-3} for the first factor and 1.3×10^{-3} for the second factor. More importantly, the results from the 100-point quadrature showed no reversals of the first and second factors. However, as discussed earlier, increasing the number of items leads to greater inaccuracy for GHEM. For example, increasing the number of items in the simulation using a sixfold repetition (i.e., 192 items) of the true factor loadings provided in Table 1 leads to ASE's of 1.5×10^{-2} and 2.1×10^{-3} for the first two factors from GHEM, compared to 1.1×10^{-3} and 1.3×10^{-3} from MCEM; both methods used 100 points.

3.2 Example 2: A Five-Factor Simulation

Increasing the number of quadrature points in GHEM can yield fairly accurate results in low dimensions with moderate numbers of items. However, achieving the necessary accuracy in higher dimensions by increasing the number of points quickly becomes impractical, because the number of points increases exponentially with the dimension. To illustrate the comparative accuracy of the methods in higher dimensions, we repeat the foregoing simulation but with a five-factor model using the true factor loadings provided in Table 2, and again set all item difficulties to zero. We applied MCEM and GHEM with 25 and 243 points, which should favor GHEM. But as Table 2 shows, the ASE's from MCEM are nearly an order of magnitude smaller for the first two factors and are substantially smaller for the remaining factors. Moreover, the MCEM estimates for the first factor are always greater in magnitude (with equality for item 23) than the second. In contrast, the GHEM estimates showed 11 reversals. Reversals of the

Table 2. Estimates and True Values for Example 2

	Factor 1			Factor 2			Factor 3			Factor 4			Factor 5		
Item	GHEM	MCEM	True	GHEM	MCEM	True	GHEM	MCEM	True	GHEM	MCEM	True	GHEM	MCEM	True
1	.58	.49	.50	−.37	−.40	−.37	.34	.36	.34	.12	.31	.31	−.23	−.15	−.27
2	.60	.50	.50	−.36	−.41	−.37	.32	.37	.34	.07	.25	.31	−.23	−.17	−.27
3	.45	.44	.50	−.52	−.43	−.37	.22	.22	.34	.27	.32	.31	.19	.32	.27
4	.49	.49	.50	−.46	−.36	−.37	.24	.24	.34	.29	.30	.31	.25	.36	.27
5	.62	.46	.50	−.26	−.44	−.37	.15	.32	.34	−.30	−.15	−.31	−.20	−.26	−.27
6	.59	.47	.50	−.20	−.41	−.37	.14	.38	.34	−.41	−.28	−.31	−.14	−.21	−.27
7	.49	.45	.50	−.28	−.36	−.37	.17	.33	.34	−.21	−.24	−.31	.31	.31	.27
8	.54	.48	.50	−.32	−.41	−.37	.15	.32	.34	−.20	−.24	−.31	.35	.34	.27
9	.44	.54	.50	−.18	−.32	−.37	−.35	−.36	−.34	.24	.29	.31	−.35	−.27	−.27
10	.43	.51	.50	−.18	−.35	−.37	−.33	−.34	−.34	.17	.24	.31	−.36	−.30	−.27
11	.38	.50	.50	−.29	−.32	−.37	−.36	−.42	−.34	.38	.25	.31	.14	.28	.27
12	.42	.53	.50	−.26	−.31	−.37	−.32	−.37	−.34	.38	.26	.31	.07	.19	.27
13	.54	.55	.50	−.08	−.37	−.37	−.45	−.29	−.34	−.34	−.38	−.31	−.16	−.30	−.27
14	.51	.55	.50	−.05	−.32	−.37	−.40	−.25	−.34	−.34	−.33	−.31	−.21	−.32	−.27
15	.51	.58	.50	−.18	−.33	−.37	−.45	−.31	−.34	−.16	−.37	−.31	.33	.28	.27
16	.50	.58	.50	−.15	−.32	−.37	−.45	−.33	−.34	−.07	−.26	−.31	.31	.27	.27
17	.35	.47	.50	.29	.33	.37	.40	.34	.34	.24	.38	.31	−.24	−.20	−.27
18	.39	.46	.50	.25	.27	.37	.43	.39	.34	.17	.35	.31	−.34	−.31	−.27
19	.27	.40	.50	.23	.38	.37	.43	.30	.34	.38	.36	.31	.17	.27	.27
20	.31	.47	.50	.20	.37	.37	.36	.24	.34	.40	.37	.31	.20	.31	.27
21	.44	.48	.50	.44	.32	.37	.33	.48	.34	−.36	−.31	−.31	−.11	−.26	−.27
22	.45	.49	.50	.40	.31	.37	.35	.49	.34	−.31	−.25	−.31	−.09	−.26	−.27
23	.31	.39	.50	.38	.39	.37	.36	.44	.34	−.18	−.33	−.31	.43	.34	.27
24	.33	.44	.50	.37	.41	.37	.27	.30	.34	−.09	−.22	−.31	.43	.35	.27
25	.30	.53	.50	.43	.39	.37	−.10	−.23	−.34	.33	.30	.31	−.34	−.30	−.27
26	.35	.57	.50	.46	.33	.37	−.12	−.20	−.34	.26	.26	.31	−.37	−.34	−.27
27	.23	.50	.50	.38	.42	.37	−.21	−.37	−.34	.44	.23	.31	.13	.17	.27
28	.24	.52	.50	.36	.41	.37	−.23	−.40	−.34	.52	.25	.31	.22	.29	.27
29	.35	.53	.50	.53	.38	.37	−.19	−.14	−.34	−.19	−.25	−.31	−.16	−.32	−.27
30	.39	.58	.50	.52	.39	.37	−.24	−.18	−.34	−.18	−.30	−.31	−.05	−.19	−.27
31	.31	.55	.50	.52	.44	.37	−.31	−.30	−.34	−.02	−.31	−.31	.29	.16	.27
32	.26	.47	.50	.52	.44	.37	−.30	−.29	−.34	−.09	−.36	−.31	.31	.19	.27
ASE	.019	.002		.02	.002		.012	.007		.018	.003		.011	.004	

second through fifth factors were more common due to the closeness of the true loadings. But, although the differences between GHEM and MCEM estimates in numbers of these reversals were not substantial (39 reversals from GHEM and 32 from MCEM), the magnitude of the reversals were more substantial from GHEM. For example, items 13 and 14 exhibited loadings of $-.08$ and $-.05$ on the second factor and $-.45$ and $-.40$ on the third factor from the GHEM estimates. None of the reversals for MCEM were of comparable magnitude. (As an indication of the computational load of MCEM, this problem takes about 6 minutes on a Gateway P5-100XL under OS2, with 20 EM iterations.)

3.3 Example 3: LSAT Section 7 Data

As a check of the performance of MCEM in practice, we fitted a two-factor FIIF model to the data from Section 7 of the Law School Admissions Test (LSAT) first used by Bock and Lieberman (1970), who fitted a one-dimensional model to it. Bock and Aitken (1981) later found that more than one factor was necessary and fitted a two-factor model using their EM method. Because this was the first empirical application of the FIIF model, this data set has become a canonical example and was later reproduced by Bock, Gibbons, and Muraki (1988).

Convergence of MCEM with 100 points was determined from Figures 1 and 2. As Figure 1 shows, the logs of the likelihood ratios, estimated by (19) and plotted as dots (the circles are discussed in Sec. 4.2), appears to have stabilized after about six iterations. Similarly, the largest changes in the parameters (Fig. 2) appear to stabilize after six EM cycles. To be conservative, we take results after 10 iterations as our MCEM estimates, which are given in Table 3 along with those from GHEM and the MINRES solution used as the initial values. In Example 2, we saw that GHEM achieved comparable accuracy with MCEM for the 2-factor, 32-item simulation by using 100 quadrature points. This is also the case for these real data. The parameter estimates are essentially identical; indeed, both methods show little change from the initial values from MINRES. This provides additional evidence, albeit in a roundabout way, for the accuracy of MCEM, including our convergence monitoring scheme, because the Gauss–Hermite approach is accurate in such a low-dimensional problem.

3.4 Example 4: The 1978 Quality of American Life Survey

To compare MCEM to GHEM in high-dimension real-data problems, we chose to examine the quality of life survey (Campbell and Converse 1980; Campbell, Converse, and Rodgers 1976) administered to a nationwide probability sample of all U.S. residents age 18 and older in both 1974 and 1978. Bock et al. (1988) applied the FIIF model

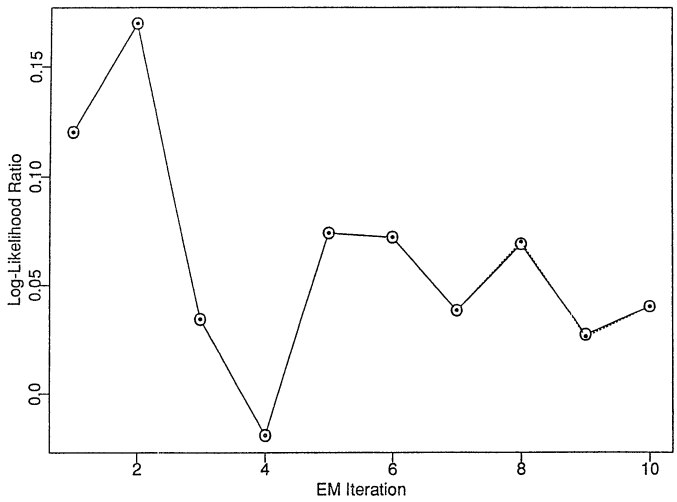

Figure 1. Example 3: Log-Likelihood Ratio Versus EM Iteration Using $\hat{r}_G$ (denoted by •) and $\hat{r}_O$ (denoted by ○).

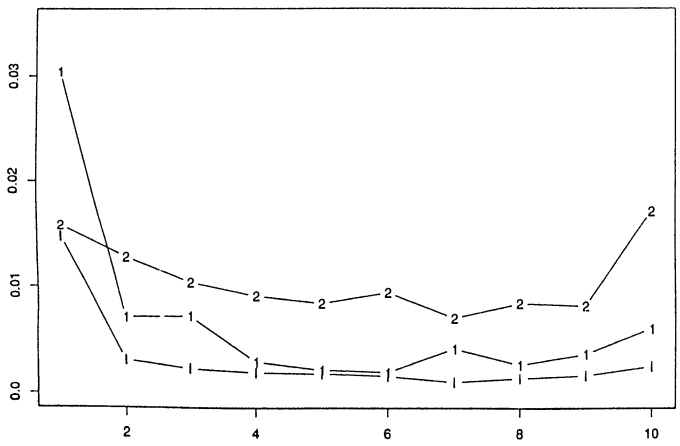

Figure 2. Example 3: Largest Parameter Changes Versus EM Iteration. Key: 1—slope 1; 2—slope 2; I—Intercept.

to the 1974 survey using GHEM. For reasons of data accessibility, we chose a subsample 2,159 heads of household from the 1978 survey to fit a five-factor FIIF model using both GHEM and MCEM. In the survey, the subjects were asked to rate their satisfaction, on a 7-point scale, with 14 aspects of their life: satisfaction with community, neighborhood, house, life in the United States, education, health, job, leisure, friends, family, standard of living, savings, life in general, and self. Following Bock et al. (1988), these ratings were dichotomized at the neutral category.

Convergence of MCEM was determined from Figures 3 and 4. Figure 3 shows that the logs of the likelihood ratios, estimated from (19), decrease to zero after nine iterations. Likewise, the largest parameter changes also stabilize after nine iterations (Fig. 4). Again, to be conservative, the converged estimates were obtained after 15 iterations and are presented along with the corresponding GHEM estimates in Table 4, which shows considerable differences, particularly for the higher-order factors. To see the impact of such differences on the grouping of the items (i.e., aspects), a standard part of factor analysis, we examined the rotated factor loadings. Varimax rotated factor loadings (see, e.g., Kaiser 1958) from MCEM and GHEM estimates are presented in Table 5.

The loadings from MCEM estimates lead to the following exploratory grouping: factor 1—leisure, friends, family, life, and self; factor 2—community, neighborhood, house, and life in the U.S.; factor 3—standard of living and savings; factor 4—education; and factor 5—health and job. In contrast, the GHEM varimax factor loadings reveal job loading essentially equally on factors 1–3, with the last factor ill-determined by any group of items. Whether the groups from the MCEM results have substantive meaning is an issue requiring external knowledge and assessment (e.g., one may argue that the fifth factor is meaningful because it reflects the importance of employer-provided health insurance to both job and health satisfaction, but one perhaps can come up with an equally or more plausible argument for grouping job with standard of living), but it is clear that one cannot rely on numerical errors for meaningful grouping.

4. EMPIRICAL INVESTIGATION OF BRIDGE SAMPLING USING FIIF MODEL

4.1 Choosing the Bridge

In Section 2.4 we discussed the use of bridge sampling for simulating likelihood ratios to monitor the convergence of MCEM. To better understand the key identity (15) underlying the bridge sampling, we reexpress $\alpha = q_0/(q_1 q_2)$ in terms of a new function q_0. Suppose that q_0 is a nonnegative function and can be normalized into a density $p_0 = q_0/c_0$. Then (15) becomes

$$r \equiv \frac{c_2}{c_1} = \frac{c_0/c_1}{c_0/c_2} = \frac{E_1\left[\frac{q_0(\omega)}{q_1(\omega)}\right]}{E_2\left[\frac{q_0(\omega)}{q_2(\omega)}\right]}. \tag{20}$$

Comparing this to (14), we immediately see why bridge sampling can provide more efficient estimators than the importance sampling estimators based on (14). Intuitively, with (14), we use draws from p_1 to go all the way to reach p_2, whereas with (20) we use draws from p_1 and p_2 each to go "halfway" and use q_0 as a connecting "bridge" and

Table 3. Estimates for Example 3

	Intercept			Slope 1			Slope 2		
Item	Init	GHEM	MCEM	Init	GHEM	MCEM	Init	GHEM	MCEM
1	1.15	1.15	1.15	.62	.62	.62	.31	.31	.31
2	.67	.67	.67	1.08	1.08	1.08	−.73	−.73	−.73
3	.97	.96	.96	.82	.82	.80	−.05	−.05	−.06
4	.30	.30	.30	.47	.48	.48	.14	.18	.16
5	1.18	1.13	1.15	.49	.44	.45	.38	.27	.30

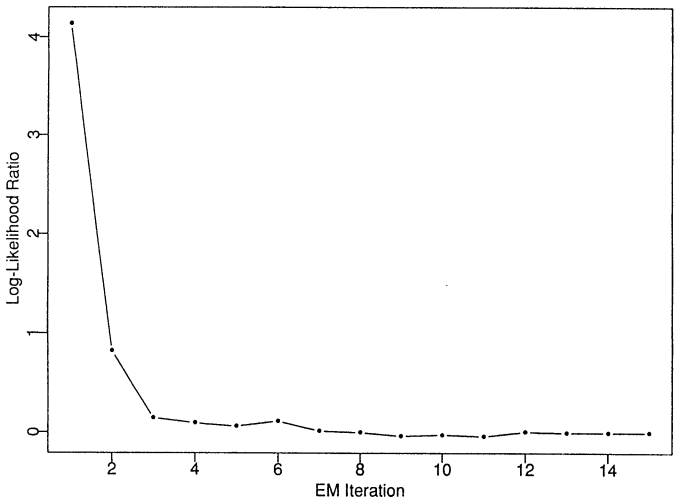

Figure 3. Example 4: Log-Likelihood Ratio Versus EM Iteration Using $\hat{r}_G$.

thus shorten the distances between the densities, which are responsible for the variabilities of the estimators. This intuition was used by Gelman and Meng (1994) to construct multibridge extensions and even the "continuous" bridge extension—the path sampling.

For given α, the corresponding bridge sampling estimator is

$$\hat{r}_\alpha = \frac{\frac{1}{n_1}\sum_{j=1}^{n_1} q_2(\omega_{1j})\alpha(\omega_{1j})}{\frac{1}{n_2}\sum_{j=1}^{n_2} q_1(\omega_{2j})\alpha(\omega_{2j})}, \tag{21}$$

where $\omega_{i1}, \ldots, \omega_{in_i}$ are (possibly dependent) draws from $p_i(\omega), i = 1, 2$. The question of interest then is how to choose a good "bridge," or the α function. Under the assumption that all draws are independent, it can be shown (e.g., see Meng and Wong 1996) that the asymptotically optimal choice of α, in the sense of minimizing the relative MSE $E(\hat{r}_\alpha - r)^2/r^2$, is given by

$$\alpha_O(\omega) = \frac{c}{s_1 r q_1 + s_2 q_2}, \quad \text{for any} \quad c \neq 0, \tag{22}$$

where $s_i = n_i/(n_1 + n_2), i = 1, 2$. Because α_O depends on r, Meng and Wong (1996) constructed an iterative sequence that monotonically (in terms of absolute differences between iterates and the limit) converges to a unique limit that achieves the optimal error. Their iterative sequence is

$$\hat{r}_O^{(t+1)} = \frac{\frac{1}{n_1}\sum_{j=1}^{n_1}\left[\frac{q_2(\omega_{1j})}{s_1\hat{r}_O^{(t)}q_1(\omega_{1j}) + s_2 q_2(\omega_{1j})}\right]}{\frac{1}{n_2}\sum_{j=1}^{n_2}\left[\frac{q_1(\omega_{2j})}{s_1\hat{r}_O^{(t)}q_1(\omega_{2j}) + s_2 q_2(\omega_{2j})}\right]}$$

$$= \frac{\frac{1}{n_1}\sum_{j=1}^{n_1}\left[\frac{l_{1j}}{s_1\hat{r}_O^{(t)} + s_2 l_{1j}}\right]}{\frac{1}{n_2}\sum_{j=1}^{n_2}\left[\frac{1}{s_1\hat{r}_O^{(t)} + s_2 l_{2j}}\right]}, \qquad \hat{r}_O^{(0)} > 0, \tag{23}$$

where $l_{ij} = q_2(\omega_{ij})/q_1(\omega_{ij}), j = 1, \ldots, n_i, i = 1, 2$ need be evaluated only once at the beginning of the iteration. The limit of $\hat{r}_O^{(t)}$ is denoted by $\hat{r}_O$, which can also be derived as a "profile maximum likelihood estimator" if one treats the problem of computing r as an estimation problem, as discussed by Geyer (1994).

Meng and Wong (1996) compared (23) to a similar iterative estimator based on importance sampling using a mixture. Specifically, considering the pooled sample $\{\omega_1, \ldots, \omega_n\} \equiv \{\omega_{ij}, j = 1, \ldots, n_i, i = 1, 2\}$ as a sample from the mixture $s_1 p_1 + s_2 p_2$, the iterative estimator based on the updating of this mixture is given by

$$\hat{r}_M^{(t+1)} = \frac{\sum_{j=1}^{n}\left[\frac{l_j}{s_1\hat{r}_M^{(t)} + s_2 l_j}\right]}{\sum_{j=1}^{n}\left[\frac{1}{s_1\hat{r}_M^{(t)} + s_2 l_j}\right]}, \hat{r}_M^{(0)} > 0, \tag{24}$$

where $l_j = q_2(\omega_j)/q_1(\omega_j), j = 1, \ldots, n$. Meng and Wong (1996) showed that this sequence has the same monotonic convergence behavior and in fact always converges to the same limit, $\hat{r}_M = \hat{r}_O$. The difference is that (23) yields a consistent estimate for r at each iteration, whereas (24) yields a consistent estimate only on convergence. For this reason, Meng and Wong (1996) conjectured that $\hat{r}_O^{(t)}$ should converge more rapidly than $\hat{r}_M^{(t)}$. Checking this conjecture is the first objective of our empirical investigation.

Meng and Wong (1996) also considered a number of noniterative choices for α, including geometric $\alpha = (\sqrt{q_1 q_2})^{-1}$, yielding

$$r_G = \frac{E_1\left(\sqrt{\frac{q_2}{q_1}}\right)}{E_2\left(\sqrt{\frac{q_1}{q_2}}\right)}, \tag{25}$$

and constant $\alpha = 1$, yielding

$$r_C = \frac{E_1(q_2)}{E_2(q_1)}. \tag{26}$$

The importance sampling formula (14) is also a special case of (15) with $\alpha = q_1^{-1}$,

$$r_S = E_1\left[\frac{q_2(\omega)}{q_1(\omega)}\right], \tag{27}$$

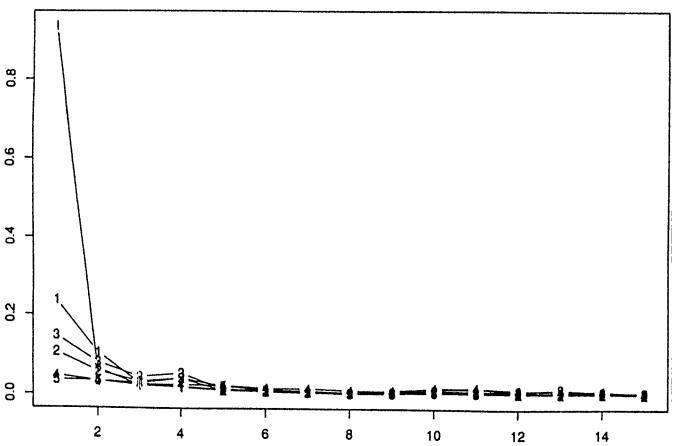

Figure 4. Example 4: Largest Parameter Changes Versus EM Iteration. Key: k —slope k, k = 1, ..., 5; I—Intercept.

Table 4. Estimates for Example 4

	Intercept		Slope 1		Slope 2		Slope 3		Slope 4		Slope 5	
Item	GHEM	MCEM	GHEM	MCEM	GHEM	MCEM	GHEM	MCEM	GHEM	MCEM	GHEM	MCEM
1	−1.10	−1.09	.93	.92	.64	.64	.41	.41	−.01	−.02	−.11	−.12
2	−1.49	−1.50	1.28	1.29	.97	.98	.44	.46	.01	.04	.00	.01
3	−1.04	−1.04	.77	.75	.44	.47	−.06	−.07	−.08	−.09	.08	.07
4	−1.05	−1.03	.68	.67	.10	.08	.23	.24	.02	.01	.00	.01
5	−.03	.07	.91	.90	.04	.04	−.27	−.27	−1.10	1.11	.38	.38
6	−.93	−1.64	.38	1.03	−.07	−.83	.19	.66	−.30	−.36	−.31	−.93
7	−1.07	−.09	.75	.51	.16	−.24	.06	.32	−.14	.09	−.46	−.28
8	−1.05	−1.00	.90	1.08	−.09	−.30	.22	.00	−.51	.06	−.54	.07
9	−1.24	−1.15	.56	.75	.04	−.17	.31	.10	−.40	.09	−.45	.06
10	−1.15	−1.08	.66	.70	−.16	−.20	.20	.07	−.09	.23	−.35	.12
11	−.87	−.91	1.30	1.62	.35	.15	−.54	−.70	−.32	.21	−.80	−.41
12	.22	.19	1.08	1.22	.41	.17	−.80	−.91	−.26	.02	−.80	−.44
13	−1.79	−1.80	1.66	1.92	−.69	−.59	.22	.01	−.31	.55	−.86	.26
14	−1.45	−1.50	.89	1.20	−.39	−.54	.23	−.02	−.56	.08	−.51	.35

where the subscript S emphasizes that it is based on draws from a single density. This choice of α corresponds to $q_0 = q_2$ in (20), which makes it clear that it does not take the advantage of the "bridge" formulation. Comparing $\hat{r}_G, \hat{r}_C, \hat{r}_S$, and $\hat{r}_O$ is the second objective of our investigations.

Our final objective is to investigate estimating $h = \int \sqrt{p_1(\omega)p_2(\omega)}\, d\omega$ via the simple identity

$$\int \sqrt{p_1 p_2}\, d\omega = \frac{E_1\left(\sqrt{\frac{q_2}{q_1}}\right)}{\sqrt{E_1\left(\frac{q_2}{q_1}\right)}}. \tag{28}$$

This is of interest here because the Hellinger distance between p_1 and p_2 is a simple function of h,

$$H(p_1, p_2) \equiv \left[\int (\sqrt{p_1} - \sqrt{p_2})^2\, d\omega\right]^{1/2} = [2(1-h)]^{1/2},$$

which was found by Meng and Wong (1996) to essentially govern the variances of bridge-sampling estimators with optimal or near-optimal choices of the bridge. Thus the Hellinger distance is a relevant "control variable" in designing investigations for comparing performance of bridge-sampling estimators. Furthermore, Meng and Wong (1996) proposed using H as a distance to decide the next density in an adaptive application of bridge sampling when we are interested in computing the ratios for a continuous range (e.g., the likelihood over a range). For this reason, they proposed using (28) to estimate h, which uses draws from only one (i.e., the previous) density. Thus the estimator $\hat{h}$ will suffer the same problem as estimators based on (14); but Meng and Wong (1996) showed that the square-root operation in (28) helps reduce the variance of $\hat{h}$. They also conjectured that $\hat{h}$ would underestimate h in general due to the larger ratios of q_2/q_1 in the denominator in (28), and thus one would overestimate H, which is desirable for constructing their adaptive estimators.

4.2 Empirical Comparisons of Various Bridge-Sampling Estimators

Our empirical study was based on fitting a one-factor FIIF model for 25 selected items from a 100-item spelling test administered to 660 undergraduate psychology students at the University of Kansas in 1987. We chose the simple one-factor model so that all relevant "golden standards" (i.e., the exact values that our simulation results will be checked against) can be obtained accurately using numer-

Table 5. Varimax Rotated Factor Loadings for Example 4

	Factor 1		Factor 2		Factor 3		Factor 4		Factor 5	
Item	GHEM	MCEM	GHEM	MCEM	GHEM	MCEM	GHEM	MCEM	GHEM	MCEM
COMMUN	.26	.16	.71	.72	.15	.13	.04	.07	.07	.18
NEIGHHD	.19	.21	.81	.81	.18	.16	.09	.08	.05	.08
HOUSE	.10	.20	.55	.53	.28	.29	.23	.21	−.06	−.03
LIFE IN US	.30	.34	.47	.41	.07	.08	.09	.10	−.17	.21
EDUCATION	.26	.18	.19	.20	.17	.17	.75	.77	.00	.05
HEALTH	.48	.25	.12	.07	.10	.10	.10	.12	.08	.82
JOB	.45	.29	.35	.17	.36	.05	.03	−.06	.00	.47
LEISURE	.66	.60	.24	.21	.24	.25	.19	.18	.03	.25
FRIENDS	.57	.50	.24	.23	.14	.15	.10	.10	.16	.21
FAMILY	.54	.55	.23	.19	.17	.14	.01	.01	−.14	.15
STAN OF LIV	.40	.40	.31	.32	.68	.69	.16	.10	−.03	.15
SAVINGS	.27	.25	.22	.20	.77	.76	.14	.16	.00	.08
LIFE	.77	.80	.20	.23	.29	.28	.09	.05	−.28	.22
SELF	.72	.72	.11	.12	.18	.17	.22	.22	−.07	.18

Table 6. Comparison of $\hat{r}_O^{(t)}$ and $\hat{r}_M^{(t)}$

	$t = 0$	$t = 1$	$t = 2$	$t = \infty$	No. Iter
		$H = .28,\ r = .08078$			
$\hat{r}_O^{(t)}$					
Mean	.13863	.08180	.08085	.08086	7.20
Var	4.43002	.00012	.00002	.00002	1.69
$\hat{r}_M^{(t)}$					
Mean	.13863	.07949	.08074	.08086	10.38
Var	4.43002	.00007	.00002	.00002	.95
		$H = .77,\ r = .56647$			
$\hat{r}_O^{(t)}$					
Mean	.61321	.59120	.58086	.58201	10.76
Var	5.93885	.02016	.01652	.01665	8.31
$\hat{r}_M^{(t)}$					
Mean	.61321	.52757	.54946	.58201	32.62
Var	5.93885	.09351	.02263	.01665	14.49
		$H = 1.11,\ r = 46.87$			
$\hat{r}_O^{(t)}$					
Mean	51.37	54.82	51.75	52.34	14.80
Var	100,166	650.20	417.97	433.97	28.35
$\hat{r}_M^{(t)}$					
Mean	51.37	43.38	42.27	52.34	96.77
Var	100,166	16,454.4	4,552.70	433.97	55.12

ical integrations. We used the PQUAD program (Wichura 1989) for computing all requisite one-dimensional integrations, with relative accuracy of 10^{-12}. We chose two sets of item parameters to be used as θ_1 and θ_2 : θ_1, consisting of the MLE's obtained by MCEM; and θ_2, consisting of all slopes equal to 1 and all intercepts equal to zero (i.e., we were computing the likelihood ratio needed for testing $\theta = \theta_2$).

As we have seen in (16), the overall likelihood ratio is the product of individual likelihood ratios based on each response patterns. To vary the Hellinger distances between p_1 and p_2, we selected three response patterns, $\mathbf{u}_i, i = 1, 2, 3$, such that the H distance between $p_1 = f(\mathbf{z}|\mathbf{u}_i, \theta_1)$ and $p_2 = f(\mathbf{z}|\mathbf{u}_i, \theta_2)$ are small, medium, and large ($H = .28, .77, 1.11$). For each selected response pattern, we used the Gibbs sampler described in Section 2.3 to make 100

Table 7. Comparison of Four Estimators

	$\hat{r}_S$	$\hat{r}_C$	$\hat{r}_G$	$\hat{r}_O$	$\hat{h}$
		$H = .28,\ r = .08078,\ h = .96$			
Mean	.080600	.081590	.080730	.080860	.961400
Var	.000137	.000063	.000022	.000020	.000900
MSE	.000137	.000063	.000022	.000020	.000912
		$H = .77,\ r = .56647,\ h = .71$			
Mean	.53733	.58586	.57963	.58201	.75833
Var	.11621	.01831	.02034	.01665	.00607
MSE	.11695	.01867	.02049	.01687	.00882
		$H = 1.11,\ r = 46.87,\ h = .38$			
Mean	32.32	55.49	54.60	52.34	.60990
Var	13,017.54	691.13	709.20	433.97	.01500
MSE	13,216.20	764.82	768.35	463.47	.06610

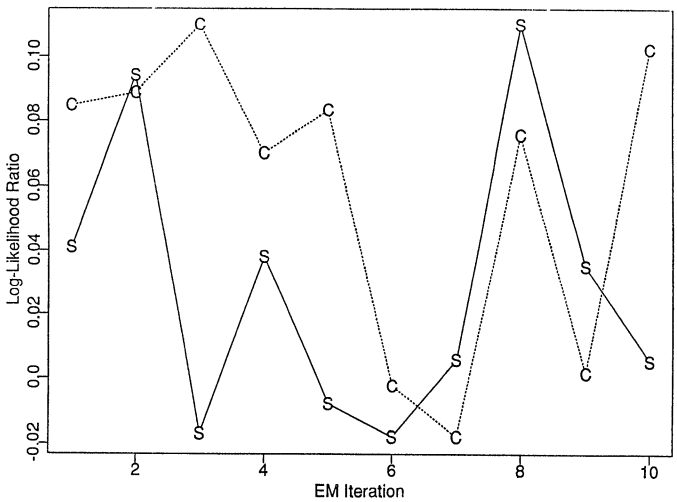

Figure 5. Example 3: Log-Likelihood Ratio Versus EM Iteration Using $\hat{r}_S$ (denoted by S) and $\hat{r}_C$ (denoted by C).

draws from each of $p_k, k = 1, 2$. To make a fair comparison with the importance sampling using draws from one density, we used the first 50 draws from each distribution to compute $\hat{r}_O, \hat{r}_M, \hat{r}_G$, and $\hat{r}_C$ and used the entire 100 draws from p_1 to compute $\hat{r}_S$ (and $\hat{h}$). We repeated this process 1,000 times, yielding 1,000 estimates for each method and each response pattern.

We first compare the two iterative procedures, $\hat{r}_O^{(t)}$ and $\hat{r}_M^{(t)}$. We constructed the starting value for both iterations, $\hat{r}^{(0)}$, from the (formal) identity

$$r = \frac{E_1[q_1^{-1}(\omega)]}{E_2[q_2^{-1}(\omega)]}, \tag{29}$$

which corresponds to choose (formally) $\alpha = [q_1(\omega)q_2(\omega)]^{-1}$ in (15). These types of estimators have been discussed in the literature (see, e.g., Gelfand and Day 1994 and Newton and Raftery 1994). This starting value was chosen because it is an extremely variable and in fact inconsistent estimator unless the support of p_k ($k = 1, 2$) has finite Lebesgue measure, which is not the case for the current setting. We purposely chose such a poor starting value to illustrate the remarkable robustness of the iteration $\hat{r}_O^{(t)}$ to the starting value. The convergence criterion of both iterations is 10^{-12}, but to prevent idiosyncratic cases in simulation, we terminated an iteration process if it exceeded 100 iterations, which happened only for $\hat{r}_M^{(t)}$.

The results of this comparison are presented in Table 6, where the means and variances are computed over the 1,000 simulations. As Table 6 shows, $\hat{r}_O^{(t)}$ and $\hat{r}_M^{(t)}$ performed nearly as well when the distance between the distributions was small (i.e., $H = .28$). However, as Meng and Wong (1996) predicted, $\hat{r}_O^{(t)}$ converged much more rapidly than $\hat{r}_M^{(t)}$ as the distances between the distributions increased. For example, when $H = 1.11$, the variance of $\hat{r}_O^{(t)}$ decreased from more than 100,000 at iteration zero to 650 at iteration 1 and to 418 at iteration 2. In contrast, by iteration 1 the variance of $\hat{r}_M^{(t)}$ decreased only to 16,454, and by iteration 2 the variance still exceeded 4,500. Accordingly and

moreover, $\hat{r}_{\mathrm{M}}^{(t)}$ takes much longer to converge, nearly seven times slower on average, *without* taking into account that the maximum number of iteration allowed was 100.

The next comparison is among $\hat{r}_{\mathrm{S}}, \hat{r}_{\mathrm{C}}, \hat{r}_{\mathrm{G}}$, and $\hat{r}_{\mathrm{O}}$; the results are presented in Table 7. The results closely parallel those of Meng and Wong's (1996) theoretical example with $p_1 = N(\mu, 1)$ and $p_2 = N(0, 1)$. As expected, $\hat{r}_{\mathrm{O}}$ dominated all of the other estimators in all cases, even when the draws from the Gibbs sampler were only approximately independent, and thus Meng and Wong's (1996) theoretical results can be taken only as a guideline. Also, $\hat{r}_{\mathrm{S}}$ is dominated by all other three estimators. In addition, $\hat{r}_{\mathrm{G}}$ performed nearly as well as $\hat{r}_{\mathrm{O}}$ when $H = .28$, and $\hat{r}_{\mathrm{C}}$ and $\hat{r}_{\mathrm{G}}$ are quite similar when $H = .77$ and $H = 1.11$. A difference from Meng and Wong's normal example is that $\hat{r}_{\mathrm{C}}$ is closer to $\hat{r}_{\mathrm{G}}$ than to $\hat{r}_{\mathrm{O}}$, whereas in their example $\hat{r}_{\mathrm{C}}$ is closer to $\hat{r}_{\mathrm{O}}$.

Another difference from Meng and Wong (1996) is that $\hat{h}$ overestimates h, especially when h is small. For example, from Table 7, when $H = 1.11$ ($h = .3838$), the mean of $\hat{h}$ is 1.6 times the true value. This seems to be due to, at least partially, the small number of the Gibbs sampler draws (i.e., 100) relative to the large H distance, a problem that could also explain the large biases in $\hat{r}_{\mathrm{S}}$. These problems are worth further investigation but do not affect our current FIIF applications, which are based on $\hat{r}_{\mathrm{G}}$ and $\hat{r}_{\mathrm{O}}$, whose performances are much more satisfactory as judged by the MSE's given in Table 7. Compared to the importance sampling estimator $\hat{r}_{\mathrm{S}}, \hat{r}_{\mathrm{G}}$ and $\hat{r}_{\mathrm{O}}$ exhibited anywhere from 5 to 30 times less MSE.

As a final illustration and comparison, we compare Figure 1 and Figure 5. Figure 1 uses both $\hat{r}_{\mathrm{G}}$ (dots) and $\hat{r}_{\mathrm{O}}$ (circles) to monitor the convergence of MCEM for the LSAT example of Section 3.3. The plots are virtually indistinguishable and allow us to reasonably access convergence after about eight iterations. In contrast, plots in Figure 5 using r_{S} or r_{C} are much more problematic—the variations around the eighth iteration are as large as those at the beginning of the iteration.

In summary, based on all of the theoretic and empirical evidences obtained so far, we recommend using $\hat{r}_{\mathrm{O}}$ in practice, with $\hat{r}_{\mathrm{G}}$ as the starting value for iterating $\hat{r}_{\mathrm{O}}^{(t)}$. This iteration requires minimum computation and converges very fast; but if iteration is not desirable (e.g., within a large simulation), then we recommend using $\hat{r}_{\mathrm{G}}$, or, with a little extra computation, $\hat{r}_{\mathrm{O}}^{(2)}$, as the estimator.

5. CONCLUDING REMARKS

Bock and Aitken's original method for fitting the FIIF model, although reliable for small numbers of items, becomes less reliable as the number of items increases. This is mainly because the predictive distributions for the individual latent abilities become more peaked as the number of items increases, leading to a "lumpy" observed-data likelihood for the model parameters, but the reliability of the fixed-point Gauss–Hermite quadrature method relies on the smoothness of the integrand. In contrast, because the Monte Carlo E step directly simulates from these individual predictive distributions, the MCEM adapts much better to the "lumpy" observed-data likelihood. However, Monte Carlo implementation is certainly not the only method for adapting to the predictive distributions. For instance, Gauss–Hermite quadrature could also be made adaptive but would require more computational effort, especially in high dimensions (e.g., needing many more quadrature points). As emphasized in Section 1, the advantages of Monte Carlo simulation are not restricted merely to point estimation; with a little extra effort, we can obtain uncertainty estimates (e.g., Meng and Rubin 1991; van Dyk, Meng, and Rubin 1995) or even a full Bayesian analysis.

Up to now, a main obstacle to effective implementation of MCEM in the FIIF model and in general has been the lack of a reliable method for determining the convergence of MCEM. This article has demonstrated how this obstacle can be effectively tackled by using bridge sampling, along the way empirically validating Meng and Wong's (1996) theoretical findings. Moreover, bridge sampling can be applied to simulate the values of the observed-data likelihoods by coupling the predictive distributions for missing data or latent variables with appropriately matched densities with known normalizing constants (e.g., a convenient importance sampling trial density). As demonstrated by Meng and Wong (1996) and here, due to the effective reduction in distance between the two densities, the likelihood values obtained from bridge sampling are often substantially more accurate than those obtained from the standard importance sampling using draws from one density.

[Received January 1995. Revised November 1995.]

REFERENCES

Albert, J. H. (1992), "Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling," *Journal of Educational Statistics*, 17, 251–269.

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Bennett, C. H. (1976), "Efficient Estimation of Free Energy Differences from Monte Carlo Data," *Journal of Computational Physics*, 22, 245–268.

Bartholomew, D. J. (1987), *Latent Variable Models and Factor Analysis*, New York: Oxford University Press.

Bock, R. D., and Aitkin, M. (1981), "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm," *Psychometrika*, 37, 29–51.

Bock, R. D., Gibbons, R., and Muraki, E. (1988), "Full-Information Item Factor Analysis," *Applied Psychological Measurement*, 12, 261–280.

Bock, R. D., and Lieberman, M. (1970), "Fitting a Response Model for Dichotomously Scored Items," *Psychometrika*, 35, 179–197.

Campbell, A., and Converse, P. E. (1980), *The Quality of American Life, 1978*, Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Campbell, A., Converse, P. E., and Rodgers, W. L. (1976), *The Quality of American Life*, New York: Russel Sage Foundation.

Cristofferson, A. (1975), "Factor Analysis of Dichotomized Variables," *Psychometrika*, 37, 29–51.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society*, Ser. B, 56, 501–514.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., and Meng, X. L. (1994), "Path Sampling for Computing Normalizing Constants: Identities and Theory," Technical Report 376, University of Chicago, Dept. of Statistics.

Gelman, A., Meng, X. L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, to appear.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–511.

Geman, S., and Geman, D. (1984), "Stochastic Simulation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, 721–741.

Geyer, C. J. (1994), "Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo," Technical Report 568, University of Minnesota, School of Statistics.

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 657–699.

Guo, S. W., and E. A. Thompson (1992), "A Monte Carlo Method for Combined Segregation and Linkage Analysis," *American Journal of Human Genetics*, 51, 1111–1126.

Gupta, A. K., and Tracy, D. S. (1976), "Recurrence Relations for the Moments of Truncated Multinormal Distribution," *Communications in Statistics, Part A—Theory and Methods*, 5, 855–865.

Harman, H. H. (1976), *Modern Factor Analysis*, Chicago: University of Chicago Press.

Irwin, M. E. (1994), "Sequential Imputation and Multilocus Linkage Analysis," Ph.D. thesis, University of Chicago, Dept. of Statistics.

Kaiser, H. F. (1958), "The Varimax Criterion for Analytic Rotation in Factor Analysis," *Psychometrika*, 23, 187–200.

Little, J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Liu, C., and Rubin, D. B. (1994), "The ECME Algorithm: A Simple Extension of EM and ECM With Fast Monotone Convergence," *Biometrika*, 81, 633–648.

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

——— (1995), "Correlation Structure and Convergence Rate of the Gibbs Sampler With Various Scans," *Journal of the Royal Statistical Society*, Ser. B, 57, 157–169.

Maxwell, A. E. (1983), "Factor Analysis," in *Encyclopedia of Statistical Sciences*, Vol. 3, eds. S. Kotz, N. L. Johnson, and C. B. Read, New York: John Wiley, pp. 2–8.

McCullough, R., and Rossi, P. E. (1991), "An Exact Likelihood Analysis of the Multinomial Probit Model," Working Paper 91-102, University of Chicago, Graduate School of Business.

Meng, X. L. (1994a), "On the Rate of Convergence of the ECM Algorithm," *The Annals of Statistics*, 22, 326–339.

——— (1994b), "Posterior Predictive p-Value," *The Annals of Statistics*, 22, 1142–1160.

Meng, X. L., and Pedlow, S. (1992), "EM: A Bibliographic Review with Missing Articles," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 24–27.

Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

——— (1992), "Recent Extensions to the EM Algorithm" (with discussion), *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Clarendon Press, Oxford, pp. 307–320.

——— (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

——— (1994), "On the Global and Component-wise Rates of Convergence of the EM Algorithm," *Linear Algebra and Its Applications* (special issue honoring Ingram Olkin), 199, 413–425.

——— (1996), "Efficient Methods for Estimating and Testing With Seemingly Unrelated Regressions in the Presence of Latent Variables and Missing Data," in *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D. A. Berry, K. M. Chalonera, and J. K. Geweke, New York: John Wiley, pp. 213–226.

Meng, X. L., and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, to appear.

Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian Inference by the Weighted Likelihood Bootstrap" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 56, 3–48.

Muthen, B. (1978), "Contributions to Factor Analysis of Dichotomized Variables," *Psychometrika*, 43, 551–560.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.

Rubin, D. B., and Thayer (1982), "EM Algorithms for ML Factor Analysis," *Psychometrika*, 47, 69–76.

Stein, M. L. (1992), "Prediction and Inference for Truncated Spatial Data," *Journal of Computational and Graphical Statistics*, 1, 91–110.

Tanner, M. A. (1991), *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, New York: Springer-Verlag.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Thurstone, L. L. (1947), *Multiple Factor Analysis*, Chicago: University of Chicago Press.

van Dyk, D. A., Meng, X. L., and Rubin, D. B. (1995), "Maximum Likelihood Estimation via the ECM Algorithm: Computing the Asymptotic Variance," *Statistica Sinica*, 5, 55–75.

Voter, A. F. (1985), "A Monte Carlo Method for Determining Free-Energy Differences and Transition State Theory Rate Constants," *J. Chemical Physics*, 82, 1890–1899.

Weeks, D. E., and Lange, K. (1989), "Trials, Tribulations, and Triumphs of the EM Algorithm in Pedigree Analysis," *IMA Journal of Mathematics Applied in Medicine & Biology*, 6, 209–232.

Wei, G. C. G., and 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.

Wichura, M. J. (1989), "An Algorithm for Patterson Gaussian Quadrature," Technical Report 257, University of Chicago, Dept. of Statistics.