cogent psychology

COGNITIVE SCIENCE & NEUROSCIENCE | RESEARCH ARTICLE

# A comparison of parameter covariance estimation methods for item response models in an expectation-maximization framework

Joshua N. Pritikin[1]*

Received: 19 August 2016
Accepted: 29 November 2016
First Published: 09 January 2017

*Corresponding author: Joshua N. Pritikin, Psychiatry Department, Virginia Commonwealth University, Richmond, VA, USA
E-mail: jpritikin@pobox.com

Reviewing editor:
Peter Walla, University of Newcastle, Australia

Additional information is available at the end of the article

**Abstract:** The Expectation-Maximization (EM) algorithm is a method for finding the maximum likelihood estimate of a model in the presence of missing data. Unfortunately, EM does not produce a parameter covariance matrix for standard errors. Both Oakes and Supplemented EM are methods for obtaining the parameter covariance matrix. SEM was discovered in 1991 and is implemented in both open-source and commercial item response model estimation software. Oakes, a more recent method discovered in 1999, had not been implemented in item response model software until now. Convergence properties, accuracy, and elapsed time of Oakes and Supplemental EM family algorithms are compared for a diverse selection IFA models. Oakes exhibits the best accuracy and elapsed time among algorithms compared. We recommend that Oakes be made available in item response model estimation software.

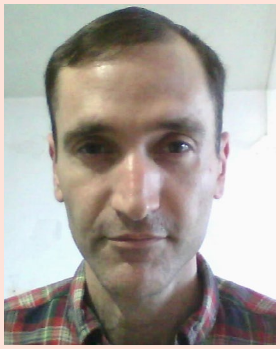
**Subjects: Multivariate Statistics; Statistical Computing; Quantitative Methods; Testing, Measurement and Assessment**

**Keywords: parameter covariance matrix; Oakes direct method; supplemented EM algorithm; item factor analysis; Monte Carlo; standard errors**

## 1. Introduction

Once a model is fit to data, it is routine practice to examine the degree of confidence we ought to have in the parameter estimates. One approximation of this information is found in the parameter covariance matrix $V$, and in summary form, as standard errors (SEs), $\sigma = \operatorname{diag}(V)^{\frac{1}{2}}$. The

## ABOUT THE AUTHOR



Joshua N. Pritikin

Joshua N. Pritikin is a postdoctoral fellow at Virginia Commonwealth University in Richmond, Virginia. Joshua is interested in developing software for the applied statistician. He has made contributions at all levels from the layout of user interface elements that are directly manipulated by applied researchers to the underlying mathematics and algorithms that implement a statistical idea.

## PUBLIC INTEREST STATEMENT

Item models are a family of a statistical models frequently used for high stakes testing (SAT, GRE, or MCAT) and for data obtained from psychological questionnaires (like personality assessments). For any statistical model, an important consideration is how precisely the model parameters are determined by the data. For these particular kind of item models, obtaining the information about parameter precision is difficult. This report runs the leading methods against each other, ranking them in terms of accuracy and performance. The method described by Oakes, which is also the simplest, outperforms all the other methods in both accuracy and performance. We recommend that Oakes be made available more widely in statistical software that deals with item models.

cogent oa

Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is a method for finding the maximum likelihood estimate (MLE, $\hat{\theta}$) of a model in the presence of missing data. For example, one EM algorithm of importance to psychologists and educators is for implementation of Item Factor Analysis (IFA) (Bock & Aitkin, 1981). Unfortunately, the parameter covariance matrix is not an immediate output of the EM algorithm. Before exploring methods to obtain the parameter covariance matrix, the EM approach will be informally outlined.

Following traditional notation, let $Y_o$ be the observed data. We want to find the MLE $\hat{\theta}$ of parameter vector $\theta$ for model $L(Y_o|\theta)$. Unfortunately, $L(Y_o|\theta)$ is intractable or cumbersome to optimize. The EM approach is to start with initial parameter vector $\theta^{t=0}$ and fill in missing data $Y_m$ as the expectation of $\{Y_m|Y_o, \theta^t\}$ (E-step). In the case of Bock and Aitkin (1981), the missing data are the examinee latent scores (as determined by item parameters). Together, the observed $Y_o$ and made-up data $Y_m$ constitute completed data $Y_c$. With the parameter vector $\theta^t$ at iteration $t$, we can use a complete data method to optimize $L(\theta|Y_c)$ and find $\theta^{t+1}$ (M-step). With an improved parameter vector $\theta^{t+1}$, the process is repeated until $\theta^t \approx \theta^{t+1} \approx \hat{\theta}$. As a memory aid, the reader may prefer to associate the $m$ in $Y_m$ with *made up* (not *missing*).

In exponential family models, the parameter covariance matrix $V$ is often estimated using the observed information matrix. The negative M-step Hessian

$$\mathcal{I}(\hat{\theta};Y_c) \approx -\frac{\partial^2 \log L(\theta|Y_c)}{\partial\theta\partial\theta} \tag{1}$$

is usually easy to evaluate but asymtotically underestimates the variability of $\mathcal{I}(\hat{\theta};Y_c)$.

A better estimate is the negative Hessian of only the observed data $Y_o$,

$$\mathcal{I}(\hat{\theta};Y_o) \approx -\frac{\partial^2 \log L(\theta|Y_o)}{\partial\theta\partial\theta}. \tag{2}$$

Usually $\mathcal{I}(\hat{\theta};Y_o)$ is difficult to evaluate; One benefit of the EM method is the ability to optimize $L(\theta|Y_o)$ efficiently without evaluation of Equation (2).

To estimate the parameter covariance matrix in an EM context, many methods have been proposed. Some methods require problem specific apparatus such as the covariance of the row-wise gradients (Mislevy, 1984) or a sandwich estimate (e.g. Louis, 1982; Yuan, Cheng, & Patton, 2013). For IFA models, the Fisher information matrix can be computed analytically. However, a sum is required over all possible patterns (Bock & Aitkin, 1981). Since such a sum is impractical for as few as 20 dichotomous items, no further consideration of this method will be given. Here we will focus on methods that are less reliant on problem specific features.

Finite differences with Richardson extrapolation has been advocated (Jamshidian & Jennrich, 2000). This method evaluates the observed data log-likelihood $\mathcal{L}(Y_o|\theta)$ at a grid of points in the $\theta$ space to approximate the Hessian. For example, for a single parameter function $f$, the Hessian can be approximated by

$$\frac{f(\theta-\epsilon) - 2f(\theta) + f(\theta+\epsilon)}{\epsilon^2} \tag{3}$$

for some small $\epsilon > 0$. For Richardson extrapolation, the perturbation distance $\epsilon$ is reduced on every iteration. Precision is enhanced by extrapolating the change in curvature between iterations (Richardson, 1911). Unfortunately, the number of points required to approximate the Hessian is $1 + r(N^2 + N)$ where $r$ is the number of iterations and $N$ is the number of parameters in vector $\theta$ (Gilbert & Varadhan, 2012). This limits the practical applicability of Richardson extrapolation to models with a modest number of parameters.

We are aware of only two algorithms that offer performance that scales linearly with the number of parameters and require little problem specific apparatus: Supplemented EM (SEM) (Meng & Rubin, 1991) and the direct method (Oakes, 1999). Although the direct method is simpler than SEM, Oakes has not been implemented in IFA software until recently (Pritikin, 2015) and has not been compared with parameter covariance estimation methods for IFA models. We describe SEM in some detail so that the reader may appreciate its relationship with Oakes.

In IFA software, SEM grew to popularity because it was superior to the methods commonly available at the time (Cai, 2008). SEM is based on the observation that the information matrix of the completed data $\mathcal{I}(\hat{\theta};Y_c)$ is the sum of the information matrices of the observed $\mathcal{I}(\hat{\theta};Y_o)$ and made-up data $\mathcal{I}(\hat{\theta};Y_m)$ (Orchard & Woodbury, 1972). With some algebraic manipulation we can rearrange the terms,

$$\mathcal{I}(\hat{\theta};Y_c) - \mathcal{I}(\hat{\theta};Y_m) = \mathcal{I}(\hat{\theta};Y_o) \tag{4}$$

$$\left[ I - \underbrace{\mathcal{I}(\hat{\theta};Y_m)\mathcal{I}^{-1}(\hat{\theta};Y_c)}_{Y_m \text{ contribution}} \right] \mathcal{I}(\hat{\theta};Y_c) = \mathcal{I}(\hat{\theta};Y_o). \tag{5}$$

Intuitively, $\mathcal{I}(\hat{\theta};Y_m)\mathcal{I}^{-1}(\hat{\theta};Y_c)$ represents the fraction of information that $Y_m$ contributes to $Y_c$ in excess of $Y_o$ (Dempster et al., 1977). One cycle of the EM algorithm can be regarded as a mapping $\theta \rightarrow M(\theta)$. In this notation, the EM algorithm is

$$\theta^{t+1} = M(\theta^t) \quad \text{for } t \in \{0, 1, \dots\}. \tag{6}$$

If $\theta^t$ converges to some point $\hat{\theta}$ and $M(\theta)$ is continuous then $\hat{\theta}$ must satisfy $\hat{\theta} \approx M(\hat{\theta})$. In the neighborhood of $\hat{\theta}$, by Taylor series expansion, $\theta^{t+1} - \hat{\theta} \approx (\theta^t - \hat{\theta})\Delta\hat{\theta}$ where $\Delta\hat{\theta}$ is the Jacobian of $M$ evaluated at the MLE $\hat{\theta}$,

$$\Delta\hat{\theta} = \left. \frac{\partial M(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}}. \tag{7}$$

Dempster et al. (1977) showed that the rate of convergence is determined by the fraction of information that $Y_m$ contributes to $Y_c$. In particular, in the neighborhood of $\hat{\theta}$,

$$\Delta\hat{\theta} \approx \mathcal{I}(\hat{\theta};Y_m)\mathcal{I}^{-1}(\hat{\theta};Y_c). \tag{8}$$

Combining Equations (5) and (8), we obtain $\mathcal{I}(\hat{\theta};Y_o) \approx \left(I - \Delta\hat{\theta}\right)\mathcal{I}(\hat{\theta};Y_c)$. Therefore, the inverse observed data parameter covariance matrix $V^{-1} \approx \left(I - \Delta\hat{\theta}\right)\mathcal{I}(\hat{\theta};Y_c)$.

The rate matrix $\Delta\hat{\theta}$ from Equation (7) can be approximated using a forward difference method. Let $d$ be the number of elements in vector $\theta$ so we can refer to it as $\theta = \{\theta_1, \dots, \theta_d\}$. Column $j$ of $\Delta\hat{\theta}$ is approximated by

$$r_{\cdot j}(\epsilon) = \frac{M(\hat{\theta}_1, \dots, \hat{\theta}_{i-1}, \hat{\theta}_j + \epsilon, \hat{\theta}_{i+1}, \dots, \hat{\theta}_d) - M(\hat{\theta})}{\epsilon}. \tag{9}$$

That is, we run 1 cycle of EM with $\theta$ set to the MLE $\hat{\theta}$ except for the $j$th parameter of $\theta$ which is set to $(\hat{\theta}_j + \epsilon)$ where $|\epsilon| > 0$ (Note that indices $i$ and $j$ are interchangeable on the diagonal). Then we subtract $M(\hat{\theta}) \approx \hat{\theta}$ from the result and divide by the scalar $\epsilon$. This amounts to numerically differentiating the EM map $M$.

Theoretically, accuracy improves as $\epsilon \rightarrow 0$. In practice, however, this is arithmetic on a computer using a floating-point representation. We cannot take $\epsilon \rightarrow 0$ but must pick a particular $|\epsilon| > 0$. The original formulation proposed to use the EM convergence history $\theta_j^t$ (where $\theta^t$ is the parameter vector $\theta$ at iteration *t*) and compute the series of columns $\{r_{\cdot j}(\theta_j^t - \hat{\theta}_j), r_{\cdot j}(\theta_j^{t+1} - \hat{\theta}_j), \dots\}$ until $r_{\cdot j}$ is "stable" from *t* to $t+1$. This procedure may initially seem appealing, but note that the history of $\theta$ is a function of the starting parameter vector $\theta^{t=0}$ and no guidance was provided about appropriate starting values. Regardless of starting values, Meng and Rubin (1991) suggested that $r_{\cdot j}$ could be declared stable if no element changed by more than the square root of the tolerance of an EM cycle. For example, if the EM tolerance for absolute change in log-likelihood is $10^{-8}$ then the SEM tolerance would be $10^{-4}$. Hence, the *j*th column of $r_{\cdot j}$ is converged when

$$|r_{ij}(\theta_j^t - \hat{\theta}_j) - r_{ij}(\theta_j^{t+1} - \hat{\theta}_j)| < \text{ tolerance } \quad \forall i \in \{1, \dots, d\} \tag{10}$$

But they remarked that the stopping criterion deserved further investigation.

SEM as originally described does not perform well in some circumstances. Its disappointing performance prompted at least two refinements. One refinement proposed a heuristic for the search of the parameter trajectory history (Tian-SEM; Tian, Cai, Thissen, & Xin, 2013) and was reinforced by a report of promising performance in unidimensional and multidimensional item response models simulation studies (Paek & Cai, 2014). The idea of Tian-SEM is that parameter estimates $\theta^t$ typically start far from the MLE $\hat{\theta}$ and approach closely only after a number of EM cycles. Starting SEM from $\theta^{t=0}$ is usually wasteful because $\Delta\hat{\theta}$ does not stabilize until $\theta^t$ with *t* close to convergence. During an EM run, the log-likelihood $\mathcal{L}$ typically changes rapidly and then slowly as the parameter values are fine tuned. The quantity $\delta^t = \exp\left(-\left|\mathcal{L}^t - \mathcal{L}^{t+1}\right|\right)$ was proposed as a standardized measure of closeness to convergence and suggested that the best opportunity for SEM is history subset $\theta^t$ corresponding to $\delta^t \in [.9, .999]$. This refinement helps in many cases. However, lingering weaknesses in Tian-SEM prompted another more drastic refinement (Agile-SEM; Pritikin, 2016). Agile-SEM will be shown to perform better than other SEM family algorithms, but not as well as the best method. A full description of Agile-SEM is lengthy and beyond the scope of this article. We include the original algorithm (MR-SEM) and these two refinements in our comparison.

Recall that the goal of SEM family algorithms is to estimate $\mathcal{I}(\hat{\theta};Y_m)$ in Equation (4). Oakes gave a remarkably direct way to obtain this quantity,

$$\mathcal{I}(\hat{\theta};Y_m) \approx \frac{\partial^2 \log L(\hat{\theta}|Y_o, Y_m)}{\partial\hat{\theta}\partial Y_m}. \tag{11}$$

This is the Jacobian of the completed data gradient of the vector $\hat{\theta}$ in $\log L(\hat{\theta}|Y_o, Y_m)$ with respect to the made up data $Y_m$ (Oakes, 1999). We are not aware of an analytic expression for this quantity for an item response model, but little additional computer programming is needed to estimate it using finite differences.

## 2. Method

### *2.1. Models*

We introduce a set of conditions designed to present a challenge to parameter covariance matrix estimators. We included underidentified models, models with bounds, and latent distribution parameters. Underidentified models do not contain enough data to uniquely identify the most likely model parameters. IFA models consist of a collection of response models, each item response model associated with a single item. To some extent, the number of possible response outcomes determines the choice of item model. The dichotomous or graded response model is suitable for items with only two possible outcomes (e.g. correct/incorrect). In contrast, the graded response or nominal model is suitable for items with more than two possible outcomes. Many other response models are available, but we focus on these three because of their enduring popularity. The definitions of these item response models are given in Appendix and detailed in the rpf package (Pritikin, 2015). The structure of Models m2pl5, m3pl15, grm20, and cyh1 will be described.

Model m2pl5 contained 5 2PL items. Slopes were 0.5, 1.4, 2.2, 3.1, and 4. Intercepts were −1.5, −0.75, 0, 0.75, and 1.5. Data were generated with a sample size of 1000 and all parameters were estimated. Model m2pl5 is not always identified at this sample size. This allowed us to examine the extent to which algorithms agreed on whether a given model was identified or not.

Model m3pl15 contained 15 3PL items. Slopes were set to 2 and items were divided into 3 groups of 5. Each group had the intercepts set as in Model m2pl5 and the lower bound parameters set to $\mathrm{logit}((1+g)^{-1})$ with $g$ as the group number (1–3). A sample size of 250 was used. For estimation, all slopes were equated to a single slope parameter. To stabilize the model, a Gaussian Bayesian prior on the lower bound (in logit units) with a standard deviation of 0.5 was used (see, Cai, Yang, & Hansen, 2011, Appendix A).

Model grm20 contained 20 graded response items with 3 outcomes. Slopes were equally spaced from 0.5 to 4. The first intercept was equally spaced from −1.5 to 1.5 every 5 items. The second intercept was 0.1 less than the first intercept. A sample size of 2,000 was used and all parameters were estimated. In the graded model, intercepts must be strictly ordered (Samejima, 1969). The placement of intercepts so close together should boost curvature in the information matrix.

The first simulation study from Cai et al. (2011) was included. Model cyh1 was a bifactor model with 2 groups of 1,000 samples each. Group 1 had 16 2PL items with the latent distribution fixed to standard Normal. Group 2 had the first 12 of the items from Group 1. All item parameters appearing in both groups were constrained equal. Data generating parameters for the items are given in Table 1. The latent distribution of Group 2 was estimated. Latent distribution generating parameters were 1, −0.5, 0, 0.5 and 0.8, 1.2, 1.5, 1, for means and variances respectively.

In addition, a 20 item 2PL model and the model from the second simulation study of Cai et al. (2011) were examined. Little additional insight was gained from these models and we do not report them here in detail. However, this work indicated that our results generalize to the nominal response model (see Appendix A).

**Table 1. Data generating parameters for Model cyh1. Group 2 did not contain items 13-16**

| Item | a1 | a2 | a3 | a4 | a5 | c |
|---|---|---|---|---|---|---|
| 1 | 1.00 | 0.80 | | | | 1.00 |
| 2 | 1.40 | 1.50 | | | | 0.25 |
| 3 | 1.70 | 1.20 | | | | −0.25 |
| 4 | 2.00 | 1.00 | | | | −1.00 |
| 5 | 1.40 | | 1.00 | | | 1.00 |
| 6 | 1.70 | | 0.80 | | | 0.25 |
| 7 | 2.00 | | 1.50 | | | −0.25 |
| 8 | 1.00 | | 1.20 | | | −1.00 |
| 9 | 1.70 | | | 1.20 | | 1.00 |
| 10 | 2.00 | | | 1.00 | | 0.25 |
| 11 | 1.00 | | | 0.80 | | −0.25 |
| 12 | 1.40 | | | 1.50 | | −1.00 |
| 13 | 2.00 | | | | 1.50 | 1.00 |
| 14 | 1.00 | | | | 1.20 | 0.25 |
| 15 | 1.40 | | | | 1.00 | −0.25 |
| 16 | 1.70 | | | | 0.80 | −1.00 |

Note: Nonzero parameters were estimated.

**Table 2. Descriptive summary of the Monte Carlo simulation studies**

| | ***#P*** | **Unidentified** | **log(*CondNum*)** | **max(\|*bias*\|)** | $\|\|bias\|\|_2$ | $\log(\|V^{-1}\|)$ |
|---|---|---|---|---|---|---|
| m2pl5 | 10 | 13 | 16.1 | 0.665 | 0.84 | 35 |
| m3pl15 | 31 | 6 | 8.5 | 0.306 | 0.55 | 90 |
| grm20 | 60 | 0 | 16.1 | 0.111 | 0.22 | 369 |
| cyh1 | 56 | 1 | 8.5 | 0.055 | 0.14 | 281 |

Notes: The first column is the number of free parameters in the model. Where the *Unidentified* column is 0, all trials were included. Trials were considered unidentified if the iteration limit was reached or the log condition number using the covariance of the gradients was greater than log(*CondNum*). *V* is the Monte Carlo parameter covariance matrix.

All item response models used a multidimensional parameterization (slope intercept form instead of discrimination difficulty). Hence, intercepts were multiplied by slopes in Models m2pl5, m3pl15, and grm20. Both MR-SEM and Tian-SEM strongly depend on the parameter convergence trajectory. Therefore, it is crucial to report optimization starting values. In general, all slopes were started at 1, intercepts at 0, means at 0, and variances at 1. For Model m3pl15, all lower bounds were started at their true value. Since the intercepts of the graded model cannot be set equal, for Model grm20, intercepts were started at 0.5 and −0.5 respectively.

### *2.2. Ground truth*

All models were subjected to 500 Monte Carlo trials to obtain the ground truth for the parameter covariance matrix. For each trial, data were generated with the rpf.sample function from the rpf package (Pritikin, 2015). Models were fit with Bock and Aitkin (1981) as implemented in the IFA module of OpenMx with EM acceleration enabled (Pritikin, 2015, Varadhan & Roland, 2008). For the multidimensional model, Cai (2010b) was used for analytic dimension reduction. The EM and M-step tolerance for relative change in log-likelihood,

$$\left|\frac{\mathcal{L}^t - \mathcal{L}^{t+1}}{\mathcal{L}^t}\right|, \tag{12}$$

were set to $10^{-9}$ and $10^{-12}$, respectively. The use of relative change removes the influence of the magnitude of $|\mathcal{L}|$ on the precision of $|\mathcal{L}|$. In models where the latent distribution was fixed, numerical integration was performed using a standard Normal prior. Single dimensional models used an equal interval quadrature of 49 points from Z score $-6$ to 6. The multidimensional model used an equal interval quadrature of 21 points from Z score $-5$ to 5. The computer used was running GNU/Linux with a 2.40 GHz Intel i7-3630QM CPU and ample RAM. Table 2 summarizes the results.

The condition number of the information matrix is the maximum singular value divided by the minimum singular value and provides a rough gauge of the stability of a solution (Luenberger & Ye, 2008, p. 239). For example, models that are amply overspecified have a condition number close to 0 whereas slightly overspecified models will have a large positive condition number. When the information matrix is not positive definite then the MLE is unstable and may be a saddle point (Luenberger & Ye, 2008, p. 190). For reference, bias is defined as $\mathbb{E}\,\theta - \hat{\theta}$ (columns 4 and 5) and the Monte Carlo parameter covariance matrix is simply the covariance of each trial's MLE $\hat{\theta}$ as the rows of data (column 6).

### *2.3. Measures of quality*

In theory, SEs approach 0 proportional to $N^{-\frac{1}{2}}$. In practice, however, each additional participant does not contribute exactly 1 unit of information. Relative difference (RD) is a way to transform SEs into comparable units across conditions,

$$RD = \frac{\text{SE} - \text{SE}_{\text{true}}}{\text{SE}_{\text{true}}}.$$

To summarize RDs for a set of parameters, the Euclidean or $l^2$-norm is used, $||RD||_2$.

SEs are an incomplete measure of information matrix estimation quality because they only reflect parameter variance. Accurate parameter covariances can be regarded as evidence that the estimation algorithm will generalize to other models. Kullback-Leibler (KL) divergence was used to assess the quality of the variance covariance matrix as a whole. For a 0 mean multivariate Normal distribution, KL divergence was defined as

$$D_{\text{KL}}(\Sigma_{true}, \Sigma) = \frac{1}{2}\left[Tr(\Sigma^{-1}\Sigma_{true}) - K - \log\left(\frac{|\Sigma_{true}|}{|\Sigma|}\right)\right]$$

where $K$ is the dimension of $\Sigma$.

### *2.4. Procedure*

We evaluated convergence properties, accuracy, and elapsed time of Oakes, MR-SEM, Tian-SEM, and Agile-SEM with 500 Monte Carlo replications. The completed data information matrix (Equation (1)) and central difference Richardson extrapolation with an initial step size of $10^{-3}$ and 2 iterations were included as low and high accuracy benchmarks, respectively. A relative EM tolerance of $10^{-11}$ was used without EM acceleration. This relative tolerance roughly corresponds to an absolute tolerance of $10^{-6}$ for the models of interest. The quantity $\mathcal{I}(\hat{\theta};Y_m)$ required by Oakes (Equation (11)), was estimated by forward difference with a step size of $10^{-5}$. Richardson extrapolation was not used so only $N+1$ evaluations of the M-step gradient were required.

Since both MR-SEM and Tian-SEM depend on the parameter convergence trajectory, EM acceleration was disabled for trials of these algorithms. Without EM acceleration, the EM iteration limit was raised to 750 from the default of 500 to protect many replications of Model cyh1 from early termination. SEM tolerance was set to the square root of the nominal absolute EM tolerance, $10^{-\frac{6}{2}}$ (Meng & Rubin, 1991, p. 907). Other EM and SEM tolerances could have been selected, but would involve a trade-off. Either accuracy would be improved at the cost of speed or vice versa. As will be seen, Oakes outperforms all SEM family algorithms in both accuracy and speed.

## 3. Results

Table 3 exhibits the percentage of models for which each algorithm converged. MR-SEM and Tian-SEM failed to converge for a substantial number of trials. For these algorithms, a failure to converge does not only squander the time spent due to SEM, but if SEM is to be reattempted then the model must be re-fit from starting values. Although Agile-SEM performs well, Oakes exhibits the best performance. Table 4 exhibits mean elapsed time and accuracy of parameter covariance matrix estimators. Oakes matches or outperforms all other algorithms in both accuracy and time, with the exception of the quick to estimate, low accuracy Mstep benchmark.

**Table 3. Percentage of trials that failed to converge by model and algorithm**

| | RE | Oakes | Agile | Tian | MR SEM |
|---|---|---|---|---|---|
| m2pl5 | 2.6 | 2.6 | 3.6 | 3.8 | 4.8 |
| m3pl15 | 1.0 | 1.0 | 1.0 | 1.0 | 1.2 |
| grm20 | 0.0 | 0.0 | 0.4 | 0.0 | 95.4 |
| cyh1 | 0.0 | 0.0 | 0.0 | 20.0 | 70.2 |

Notes: Failure was due to either iteration limit or a non-positive definite covariance matrix. Since some trials were genuinely unidentified, these trials failed to converge for all algorithms. Compare with the *unidentified* column in Table 2.

**Table 4. Mean elapsed time and accuracy of parameter covariance matrix estimators**

| | RE | Oakes | Agile | Tian | MR SEM | M-step |
|---|---|---|---|---|---|---|
| m2pl5 | | | | | | |
| Seconds | 0.015 | 0.015 | 0.019 | 0.035 | 0.08 | 0.012 |
| $\log(D_{KL})$ | 3.225 | 3.225 | 3.232 | 4.537 | 4.364 | 4.532 |
| $\|RD\|_2$ | 1.428 | 1.402 | 1.369 | 1.758 | 1.685 | 1.751 |
| m3pl15 | | | | | | |
| Seconds | 0.275 | 0.056 | 0.111 | 0.094 | 0.177 | 0.048 |
| $\log(D_{KL})$ | 11.893 | 11.893 | 11.893 | 12.014 | 11.944 | 12.014 |
| $\|RD\|_2$ | 5.273 | 3.408 | 5.271 | 4.772 | 5.07 | 4.771 |
| grm20 | | | | | | |
| Seconds | 8.039 | 0.143 | 0.478 | 0.515 | 0.247 | 0.026 |
| $\log(D_{KL})$ | 0.862 | 0.862 | 0.899 | 1.326 | 2.159 | 2.15 |
| $\|RD\|_2$ | 0.675 | 0.675 | 0.687 | 0.859 | 1.55 | 1.532 |
| cyh1 | | | | | | |
| Seconds | 46.358 | 0.829 | 2.953 | 9.657 | 12.18 | 0.072 |
| $\log(D_{KL})$ | 1.406 | 1.395 | 1.395 | 1.482 | 1.626 | 4.919 |
| $\|RD\|_2$ | 1.286 | 1.243 | 1.26 | 2.028 | 4.216 | 3.772 |

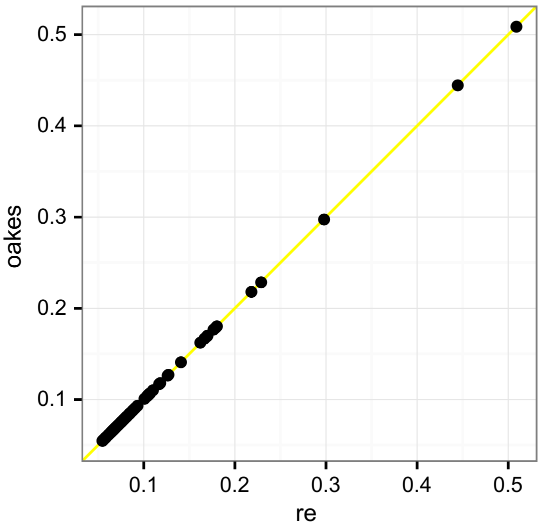
Notes: RE is central difference with Richardson extrapolation and Mstep is the completed data information matrix (Equation (1)). Since unconveraged trials were excluded, the performance of MR-SEM and Tian are shown in a most positive light. The scales of $D_{KL}$ and $\|RD\|_2$ are model specific and should not be compared between models.

## 4. Application

A subset of data from the 2009 Program for International Student Assessment (Bryer, 2012, Organisation for Economic Co-operation and Development, 2009) were used to illustrate the performance of Oakes. Responses to 35 mathematics items by students in the United States were analyzed. A total of 2,981 examinees were represented, but non-missing responses per item ranged from 1,040 to 1,618. The items were scored in 2 and 3 outcomes. Item calibration used the graded response model, consisting of 73 parameters. IFA Model Builder for OpenMx (Pritikin, 2016) was used to create the analysis script.

Both Oakes and Richardson extrapolation were used to estimate standard errors of the item parameters. As before, Richardson extrapolation used central difference with an initial step size of $10^{-3}$

**Figure 1. Scatterplot of Oakes vs. Richardson extrapolation derived standard errors.**



and 2 iterations, and Oakes used forward difference with a step size of $10^{-5}$. The EM algorithm converged with a maximum absolute gradient of $1.25 \times 10^{-2}$. For both algorithms, condition number of the information matrix was estimated at about 28. The maximum absolute difference between SE estimates was $1.31 \times 10^{-3}$. The SEs are plotted in Figure 1. Richardson extrapolation took 20.08 s and Oakes took 0.17 s.

## 5. Discussion and conclusion

We compared the convergence properties, accuracy, and elapsed time of Oakes and Supplemental EM family algorithms for a diverse selection IFA models. Oakes exhibited superior accuracy and speed. In the present article, only four models were examined. More research is needed to firmly establish whether the superior accuracy of Oakes generalizes. However, we argue that the evidence is already persuasive. When an algorithm is implemented optimally according to theoretical considerations, the deciding factor between algorithms may be the parsimony of the theory. By virtue of its theoretical simplicity, we suggest that the accuracy of Oakes cannot be surpassed by other numerical approaches.

Although SEs are a useful tool, they are not the most accurate way to assess the variability of estimated parameters. If any parameters are close to a boundary of the feasible set then likelihood-based confidence intervals should be used instead (e.g. Pek & Wu, 2015). Likelihood-based confidence intervals are comparatively slow to compute, but offer higher accuracy than a Wald test and are well supported by OpenMx.

Complete source code for all algorithms discussed is part of the OpenMx source distribution available from http://openmx.psyc.virginia.edu/. The OpenMx website additionally contains documentation and user support to assist users in analysis of their own data using item response models. OpenMx is a package for the R statistical programming environment (R Core Team, 2014).

**Funding**
The author received no direct funding for this research.

**Author details**
Joshua N. Pritikin[1]
E-mail: jpritikin@pobox.com
[1] Psychiatry Department, Virginia Commonwealth University, Richmond, VA, USA.

**Citation information**
Cite this article as: A comparison of parameter covariance estimation methods for item response models in an expectation-maximization framework, Joshua N. Pritikin, *Cogent Psychology* (2017), 4: 1279435.

**References**

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

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

Bryer, J. (2012). *PISA: Programme for International Student Assessment*. R package version 1.0. Retrieved from http://jason.bryer.org/pisa

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

Cai, L. (2010b). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581–612.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221–248.

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

Gilbert, P., & Varadhan, R. (2012). *numDeriv: Accurate numerical derivatives*. R package version. (pp. 1). Retrieved from http://CRAN.R-project.org/package=numDeriv

Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society B (Statistical Methodology), 62*, 257–270.

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

Luenberger, D. G., & Ye, Y. (2008). *Linear and nonlinear programming*. New York, NY: Springer-Verlag.

Meng, X. L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*, 899–909.

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

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society B (Statistical Methodology), 61*, 479–482.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 697–715.

Organisation for Economic Co-operation and Development. (2009). *Programme for International Student Assessment (PISA)*. http://www.pisa.oecd.org/

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling.

*Educational and Psychological Measurement, 74*, 58–76.

Pek, J., & Wu, H. (2015). Profile likelihood-based confidence intervals and regions for structural equation models. *Psychometrika, 80*, 1123–1145.

Pritikin, J. N. (2015). *rpf: Response probability functions*. R package version 0.51. Retrieved from https://CRAN.R-project.org/package=rpf

Pritikin, J. N. (2016). *A computational note on the application of the Supplemented EM algorithm to item response models.* arXiv preprint arXiv:160500860.

Pritikin, J. N., Hunter, M. D., & Boker, S. M. (2015). Modular open-source software for Item Factor Analysis. *Educational and Psychological Measurement, 75*, 458–474.

Pritikin, J. N. & Schmidt, K. M. (2016). Model builder for Item Factor Analysis with OpenMx. *R Journal, 8*(1), 182–203.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Richardson, L. F. (1911). The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character, 210*, 307–357.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(1), 1–97. doi:10.1007/BF03372160

Thissen, D., Cai, L., & Bock, R. D. (2010). *The nominal categories item response model* (pp. 43–75). Routledge.

Tian, W., Cai, L., Thissen, D., & Xin, T. (2013). Numerical differentiation methods for computing error covariance matrices in Item Response Theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement, 73*, 412–439.

Varadhan, R., & Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics, 35*, 335–353.

Yuan, K. H., Cheng, Y., & Patton, J. (2013). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, *79*, 232–254. doi:10.1007/s11336-013-9334-4

## Appendix A

### Item models

IFA models involve a set of response probability functions to appropriately model the ordinal data. The response models used in the present article are defined here. The logistic function,

$$\text{logistic}(l) \equiv \text{logit}^{-1}(l) \equiv \frac{1}{1+\exp(-l)}$$

is the basis of the response functions considered here. Due to the limits of IEEE 754 double-precision binary floating-point, the maximum absolute logit was set to 35. That is, $|l| > 35$ was clamped to $|35|$.

### *A.1. Dichotomous model*

The dichotomous response probability can model items when there are exactly two possible outcomes. It is defined as,

$$\Pr(\text{ pick} = 0|\boldsymbol{a}, c, g, \boldsymbol{\tau}) = 1 - \Pr(\text{ pick} = 1|\boldsymbol{a}, c, g, \boldsymbol{\tau})$$

$$\Pr(\text{ pick} = 1|\boldsymbol{a}, c, g, \boldsymbol{\tau}) = \text{logit}^{-1}(g) + (1 - \text{logit}^{-1}(g))\frac{1}{1+\exp(-(\boldsymbol{a}\boldsymbol{\tau}+c))}$$

where $\boldsymbol{a}$ is the slope, $c$ is the intercept, $g$ is the pseudo-guessing lower asymptote expressed in logit units, and $\boldsymbol{\tau}$ is the latent ability of the examinee (Birnbaum, 1968). A *#PL* naming shorthand has developed to refer to versions of the dichotomous model with different numbers of free parameters. Model *n*PL refers to the model obtained by freeing the first *n* of parameters *b*, *a*, and *g*.

### *A.2. Graded response model*

The graded response model is a response probability function for two or more outcomes (Cai, 2010a; Samejima, 1969). For outcomes *k* in 0 to *K*, slope vector $\boldsymbol{a}$, intercept vector $\boldsymbol{c}$, and latent ability vector $\boldsymbol{\tau}$, it is defined as,

$$\Pr(\text{pick} = 0|\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{\tau}) = 1 - \Pr(\text{pick} = 1|\boldsymbol{a}, c_1, \boldsymbol{\tau})$$

$$\Pr(\text{pick} = k|\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{\tau}) = \frac{1}{1+\exp(-(\boldsymbol{a}\boldsymbol{\tau}+c_k))} - \frac{1}{1+\exp(-(\boldsymbol{a}\boldsymbol{\tau}+c_{k+1}))}$$

$$\Pr(\text{pick} = K|\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{\tau}) = \frac{1}{1+\exp(-(\boldsymbol{a}\boldsymbol{\tau}+c_K))}.$$

### *A.3. Nominal model*

The nominal model is a response probability function for three or more outcomes (e.g. Thissen, Cai, & Bock, 2010). It can be defined as,

$$
\begin{aligned}
\boldsymbol{a} &= T_a \boldsymbol{\alpha} \\
\boldsymbol{c} &= T_c \boldsymbol{\gamma} \\
\Pr(\text{pick} = k | \boldsymbol{s}, a_k, c_k, \boldsymbol{\tau}) &= C \frac{1}{1 + \exp(-(\boldsymbol{s}\boldsymbol{\tau} a_k + c_k))}
\end{aligned}
$$

where $a_k$ and $c_k$ are the result of multiplying two vectors of free parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ by fixed matrices $T_a$ and $T_c$, respectively; $a_0$ and $c_0$ are fixed to 0 for identification; and $C$ is a normalizing constant to ensure that $\sum_k \Pr(\text{ pick} = k) = 1$.