Navigating the bowels of Marginal Maximum Likelihood: An expectation-maximization Item Response Theory estimation algorithm

Joshua N. Pritikin

Department of Psychology
University of Virginia

01 Nov 2012

Item Response Theory

Item Response Theory (IRT) helps us understand what items on a test tell us about persons, and conversely, what persons tell us about items.

IRT is an advance over Classical Test Theory because it lets us:

- construct optimal tests for specific applications
- evolve the assessment instrument over time while maintaining a stable measurement scale
- compare the ability distributions of groups (Mislevy, 1993)

IRT was conceived in the 1950s, but there was no practical way to estimate item parameters.

Birnbaum's approach

Birnbaum (1968) devised the first practical estimation algorithm known as Joint Maximum Likelihood (JML).

How does it work? JML alternates between:

- person parameters are fixed, estimate item parameters
- item parameters are fixed, estimate person parameters

Birnbaum’s approach

Advantages:

- Very general
- Usually converges

Disadvantages:

- Time proportional to # of items AND persons
- Cannot use the likelihood ratio test
- May converge to wrong estimates (Neyman & Scott, 1948)

Andersen's approach

Andersen (1972) devised an estimation algorithm known as Conditional Maximum Likelihood (CML).

Advantages:

- Avoids the Neyman-Scott Paradox

Disadvantages:

- Restricted to Rasch models
- Uses the raw score as a sufficient statistic
- Difficult to estimate a large # of items

Bock & Aitkin's approach

Bock and Aitkin (1981) devised an estimation algorithm known as Marginal Maximum Likelihood (MML). MML is an expectation-maximization (E-M) algorithm.

E-M is a general iterative method for finding maximum likelihood parameter estimates. An E-M iteration consists of two steps:

1. Create expectations for the likelihood function using provisional parameter estimates.
2. Adjust parameter estimates to maximize the likelihood.

Bock & Aitkin's approach

Advantages:

- Avoids the Neyman-Scott Paradox
- Can use likelihood ratio test
- General, not limited to Rasch models
- Linear in # of items; unlimited # of persons

Disadvantages:

- Exponential in # of person dimensions
- "An effective algorithm for MML estimation is difficult to program. Effective computer programs must be both computationally and numerically sophisticated" (Embretson & Reise, 2000, p. 214)

Schilling & Bock's refinement

Schilling and Bock (2005) proposed two refinements for Bock and Aitkin (1981) to address weaknesses in the original algorithm.

Advantages:

- Up to 15 person ability dimensions are feasible
- Even faster and more accurate convergence

Disadvantages:

- Exponential in # of person dimensions

More recent work

A recent review discussed the pros and cons of newer algorithms (Wirth & Edwards, 2007):

- Variations on Markov Chain Monte Carlo
- Underlying Bivariate Normal (Jöreskog & Moustaki, 2001)

However, Schilling and Bock (2005) remains a leading approach.

Through the bowels

How does Bock and Aitkin (1981) work?

WARNING:

Avert your eyes if you have a weak stomach.

Through the bowels

http://www.metro.co.uk/lifestyle/877834-go-inside-natures-giants-with-book-to-accompany-channel-4-series

Overview

E-M is similar to optimization by simulation.

For Bock and Aitkin (1981), in the E-step we combine provisional item estimates ξ_n with measured response frequencies $\mathbf{U}$ to obtain weights w_n. In the M-step, we optimize item parameters ξ_{n+1} using the weighted w_n likelihood.

$$\begin{aligned}\text{E-step: } & w_n = \mathrm{P}(\xi_n)\mathbf{U} \\ \text{M-step: } & \text{Maximize } \xi_{n+1} \text{ for } w_n \mathrm{P}(\xi_{n+1})\end{aligned}$$

Why not use $\mathbf{U}$ directly? The information is dispersed in the wrong shape.

A worked example

2PL item model is $\mathrm{P}(\theta) = \frac{1}{1+e^{-a(\theta-b)}}$

Take known true parameters:

a	b
0.73	0.18
2.22	0.33

Goal: Recover item parameters from simulated data.

1. E-step (marginalize person ability)
2. E-step (expected # of outcomes by quadrature)
3. M-step (Newton-Raphson or similar)
4. Repeat until convergence criteria are met

Response patterns are sparse

Item responses ($N = 4$):

item-A	item-B
1	2
2	2
1	2
1	1

Response pattern frequency distribution:

pattern	count
1,1	1
2,1	0
1,2	2
2,2	1

E-step, marginalize person ability

Holding provisional item parameters fixed,

	-4	-3.11	-2.22	-1.33	-0.44	0.44	1.33	2.22	3.11	4	total
1,1	0.86	0.75	0.60	0.41	0.24	0.12	0.05	0.02	0.01	0.00	0.204
2,1	0.01	0.02	0.03	0.04	0.05	0.05	0.04	0.03	0.02	0.01	0.045
1,2	0.13	0.22	0.36	0.50	0.59	0.59	0.50	0.36	0.22	0.13	0.547
2,2	0.00	0.01	0.02	0.05	0.12	0.24	0.41	0.60	0.75	0.86	0.204

Marginalize person ability, providing an estimate of unconditional response pattern probability. Observe that all columns sum to 1.

Assumption: Person ability is normally distributed.

E-step, marginalize person ability

Estimate the probability of a response pattern, conditional on person ability.

Response pattern: 1,2 Quadrature point: -0.44

	a	b	P(=1)	P(=2)
item-A	0.79	1.57	0.83	0.17
item-B	0.79	-1.57	0.29	0.71

$$\text{P(item-A=2}| - 0.44) = \frac{1}{1+e^{-a(\theta-b)}} = \frac{1}{1+e^{-0.79(-0.44-1.57)}} = .17$$

$$\underbrace{(1-.17)(0.71)}_{\text{P(A=1)P(B=2)}} = (0.83)(0.71) = 0.59$$

E-step, marginalize person ability

Holding provisional item parameters fixed,

	-4	-3.11	-2.22	-1.33	-0.44	0.44	1.33	2.22	3.11	4	total
1,1	0.86	0.75	0.60	0.41	0.24	0.12	0.05	0.02	0.01	0.00	0.204
2,1	0.01	0.02	0.03	0.04	0.05	0.05	0.04	0.03	0.02	0.01	0.045
1,2	0.13	0.22	0.36	0.50	0.59	0.59	0.50	0.36	0.22	0.13	0.547
2,2	0.00	0.01	0.02	0.05	0.12	0.24	0.41	0.60	0.75	0.86	0.204

Marginalize person ability, providing an estimate of unconditional response pattern probability. Observe that all columns sum to 1.

Assumption: Person ability is normally distributed.

Digression: Gauss-Hermite quadrature

G-H quadrature is an excellent approximation of integrals of the form:

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i)$$

where n is the number of sample points to use for the approximation, x_i are the roots of the Hermite polynomial, and the associated weights w_i are given by:

$$w_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 \left[H_{n-1}(x_i)\right]^2}$$

Gauss-Hermite quadrature

The current implementation uses a fixed 10-point quadrature of the Normal density.

x	area
-4.00	0.000119
-3.11	0.002805
-2.22	0.030020
-1.33	0.145800
-0.44	0.321300
0.44	0.321300
1.33	0.145800
2.22	0.030020
3.11	0.002805
4.00	0.000119

Schilling and Bock (2005) devised a more accurate adaptive quadrature approach.

E-step, marginalize person ability

Holding provisional item parameters fixed,

	-4	-3.11	-2.22	-1.33	-0.44	0.44	1.33	2.22	3.11	4	total
1,1	0.86	0.75	0.60	0.41	0.24	0.12	0.05	0.02	0.01	0.00	0.204
2,1	0.01	0.02	0.03	0.04	0.05	0.05	0.04	0.03	0.02	0.01	0.045
1,2	0.13	0.22	0.36	0.50	0.59	0.59	0.50	0.36	0.22	0.13	0.547
2,2	0.00	0.01	0.02	0.05	0.12	0.24	0.41	0.60	0.75	0.86	0.204

Marginalize person ability, providing an estimate of unconditional response pattern probability. Observe that all columns sum to 1.

Assumption: Person ability is normally distributed.

E-step, marginalize person ability

Estimate the probability of a response pattern, unconditional on person ability.

	-4	-3.11	-2.22	-1.33	-0.44	0.44	1.33	2.22	3.11	4
1,2	0.125	0.222	0.356	0.497	0.590	0.590	0.497	0.356	0.222	0.125
G-H	0.000	0.003	0.030	0.146	0.321	0.321	0.146	0.030	0.003	0.000
	0.000	0.001	0.011	0.072	0.190	0.190	0.072	0.011	0.001	0.000

Sum the product of the w_q quadrature weights and the probabilities of outcomes o_i for each item i in a given response pattern:

$$\sum_q \mathrm{P}(o_i) w_q = 0.547$$

Assumption: Person ability is normally distributed.

E-step, marginalize person ability

Holding provisional item parameters fixed,

	-4	-3.11	-2.22	-1.33	-0.44	0.44	1.33	2.22	3.11	4	total
1,1	0.86	0.75	0.60	0.41	0.24	0.12	0.05	0.02	0.01	0.00	0.204
2,1	0.01	0.02	0.03	0.04	0.05	0.05	0.04	0.03	0.02	0.01	0.045
1,2	0.13	0.22	0.36	0.50	0.59	0.59	0.50	0.36	0.22	0.13	0.547
2,2	0.00	0.01	0.02	0.05	0.12	0.24	0.41	0.60	0.75	0.86	0.204

Marginalize person ability, providing an estimate of unconditional response pattern probability. Observe that all columns sum to 1.

Assumption: Person ability is normally distributed.

E-step, expected % of outcomes by quadrature

Find expected response pattern frequency, conditional on quadrature k.

$$E(\text{pattern}_k) = \text{pattern}_{\text{freq}} \left[\frac{P(\text{pattern}_k)}{P(\text{pattern})} \right]$$

	Freq	$P(\text{pattern}_k)$	$P(\text{pattern})$	$E(\text{pattern}_k)$
1,1	1	0.2416	0.2039	1.1845
2,1	0	0.0488	0.0452	0.0000
1,2	2	0.5904	0.5470	2.1587
2,2	1	0.1192	0.2039	0.5846

E-step, expected % of outcomes by quadrature

Assumption: Items are independent after controlling for person ability.

	E(pattern$_k$)	item-A/1	item-A/2	item-B/1	item-B/2
1,1	1.1845	1.1845	0	1.1845	0
2,1	0	0	0	0	0
1,2	2.1587	2.1587	0	0	2.1587
2,2	0.5846	0	0.5846	0	0.5846
total		3.3432	0.5846	1.1845	2.7433

and scale by quadrature area at -0.44.

$$\begin{pmatrix} 3.34 \\ 0.58 \\ 1.18 \\ 2.74 \end{pmatrix} 0.32 = \begin{pmatrix} 1.07 \\ 0.19 \\ 0.38 \\ 0.88 \end{pmatrix}$$

E-step, expected % of outcomes by quadrature

Unnormalized proportion of expected outcomes, conditional on person ability (quadrature).

	item-A/1	item-A/2	item-B/1	item-B/2
-4	0.0006	0.0000	0.0005	0.0001
-3.11	0.0127	0.0001	0.0104	0.0023
-2.22	0.1270	0.0026	0.0879	0.0416
-1.33	0.5597	0.0354	0.2946	0.3005
-0.44	1.0742	0.1878	0.3806	0.8814
0.44	0.8814	0.3806	0.1878	1.0742
1.33	0.3005	0.2946	0.0354	0.5597
2.22	0.0416	0.0879	0.0026	0.1270
3.11	0.0023	0.0104	0.0001	0.0127
4	0.0001	0.0005	0.0000	0.0006

Assumption: Items are independent after controlling for person ability.

M-step

Items i are independent so we can optimize them separately. The item response probability function P is weighted by the expected # of responses.

$$\mathcal{L}(i) = \sum_k \sum_q \mathrm{expect}_{i,k|q} * \log \mathrm{P}(i, \mathrm{outcome} = k \,|\, \mathrm{quadrature} = q) + \underbrace{\mathrm{lognormal}(a_i, 0, .5)}_{\mathrm{prior}}$$

where a_i is the 2PL discrimination parameter for item i.

Iterate until convergence

Schilling and Bock (2005) suggested that it is necessary to scale and center item parameters every cycle.

As a shortcut, the current implementation does not have convergence criteria. It simply iterates the E-M steps 4 times.

Once the model is fit, there are a variety of fit statistics:

- Are all item categories utilized?
- Person ability SE function
- Person-item map
- Person-item fit
- Item-person fit

Implementation quality

Rizopoulos (2006) described `ltm`, an `R` package that implements MML/EM.

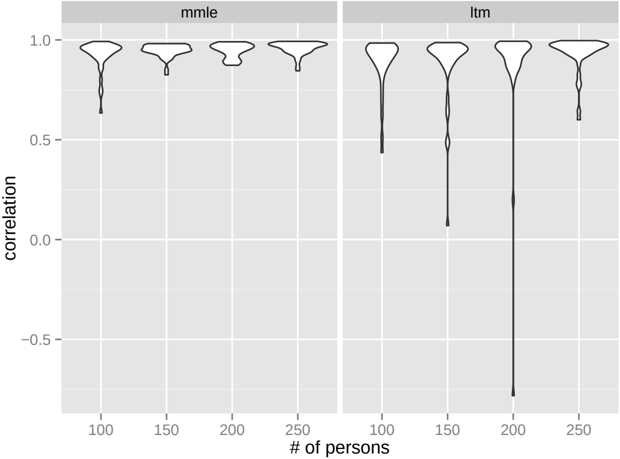
Google scholar reported 152 citations as of 2012 Oct 22.

How does `ltm` compare against our implementation ("mmle")?

- 4 Generalized Partial Credit Model items with 3 categories each
- 40 trials each for 100, 150, 200, and 250 persons

How well do we recover known true parameters from simulated data?

Results


mmle
ltm
1.0
0.5
0.0
−0.5
correlation
100
150
200
250
100
150
200
250
# of persons

Discussion

http://www.flickr.com/photos/mariusdalseg/7838988502/

You have conquered the bowels of MML/EM. Congratulations.

Future work

- Implement in C as an OpenMx objective function
- Convergence criteria, not just a fixed number of cycles
- Explore performance on ill-behaved data (with assumption violations)
- Incorporate recent refinements (Schilling & Bock, 2005)
- Compare performance against IRTPRO
- Documentation

Thank you

OpenMx development team, Mike Hunter, Karen Schmidt, Timo von Oertzen

Supported by an application (not yet funded) to NSF's Graduate Research Fellowship.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, *34*, 42–54.

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.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Lawrence Erlbaum.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347–387.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–39). New Jersey: Lawrence Erlbaum Hillsdale.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from `http://www.jstatsoft.org/v17/i05/`

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*(3), 533–555.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58.