IRT Parameter Estimation using the EM Algorithm

Brad Hanson

October 6, 1998 (revised 9/27/2000)

The EM (Expectation-Maximization) algorithm is a method for computing maximum likelihood and Bayes modal parameter estimates in situations where some data are missing (Dempster, Laird, and Rubin, 1977; McLachlan and Krishnan, 1997). The EM algorithm gives parameter estimates that maximize the likelihood of the *observed* data using computations that involve the likelihood of the *complete* data (where the complete data is the observed data plus the missing data). The EM algorithm is an iterative algorithm that involves two steps at each iteration: the E (expectation) step, and the the M (maximization) step. In the E step the expected values of the complete data sufficient statistics for the parameters are computed by averaging over the conditional distribution of the missing data given the observed data and provisional values of the parameters. In the M step parameter estimates that maximize the complete data likelihood are computed using the expected complete data sufficient statistics computed in the E step. The E and M steps are repeated until the parameter estimates converge. The EM algorithm can result in significantly simplified computation compared to trying to find parameter estimates that maximize the observed data likelihood directly. This simplification will occur when computing maximum likelihood parameter estimates using the complete data likelihood is relatively simple (i.e., if the missing data were known the parameter estimates would be simple to compute).

This paper describes computing maximum likelihood estimates of parameters in IRT models for dichotomous items using the EM algorithm. Brief descriptions of how to use the EM algorithm to compute Bayes modal parameter estimates and maximum likelihood parameter estimates for polytomous IRT models are given in the final section.

Data

Observed Data. The observed data are the responses of a sample of N examinees to J dichotomous items. The observed data are contained in a $N \times J$ matrix $\mathbf{Y}$, where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)^t$, $\mathbf{y}_i$ is a vector given by $(y_{i1}, y_{i2}, \ldots, y_{iJ})$, and y_{ij} is one if examinee i answered item j correctly, and zero if examinee i answered item j incorrectly.

Missing Data. The missing data are values of an unobserved latent variable for each examinee. The missing data are $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$, where θ_i is the value of the latent variable for examinee i. The possible values of θ_i can be real numbers (latent trait models) or categories (latent class models).

Complete Data. The complete data are the observed data plus the missing data for each examinee. The complete data are $[(\mathbf{y}_1, \theta_1), (\mathbf{y}_2, \theta_2), \ldots, (\mathbf{y}_N, \theta_N)]$.

Model

In latent trait models the latent variable is considered to be continuous (a real number that can take on any value). In this paper the latent variable is taken to be discrete in both latent trait and latent class models, and estimation procedures are derived based on the discrete latent variable. This results in exactly the same algorithm as is obtained by deriving estimation procedures based on a continuous latent variable and then implementing approximations of those procedures with a discrete version of the continuous latent variable using numerical integration (e.g., Bock and Aitken, 1981; Muraki, 1992). In this paper the specification of a discrete latent variable is done in the model itself, rather than as an approximation to a continuous latent variable for the purposes of numerical integration. This results in a more straight forward description of the EM algorithm.

It is assumed the latent variable takes on K known discrete values $q_k, k = 1, \ldots, K$, with associated unknown probabilities $\pi_k, k = 1, \ldots, K$. With this assumption the latent variable has a multinomial distribution with probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$. The notational convention used in this paper is that q_k, $k = 1, 2, \ldots, K$, are the K possible values of the latent variable, whereas θ_i is the unspecified value of the latent variable for examinee i which can equal any of the q_k.

The probability of observing item responses $\mathbf{y} = (y_1, y_2, \ldots, y_J)$ for a randomly sampled examinee from a population with a latent variable distribution given by the probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_J)$ is

$$f(\mathbf{y} \mid \boldsymbol{\Delta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} f(\mathbf{y}, q_k \mid \boldsymbol{\Delta}, \pi_k)$$

$$= \sum_{k=1}^{K} f(\mathbf{y} \mid q_k, \mathbf{\Delta})\pi_k \,, \tag{1}$$

where $\mathbf{\Delta}$ represents the item parameters which determine the probability of a particular set of item responses occurring given a fixed value of the latent variable, and $f(\mathbf{y} \mid q_k, \mathbf{\Delta})$ is the conditional probability distribution of the item responses for examinees with a value of the latent variable equal to q_k. Note that the probability given in Equation 1 is the marginal probability of the observed data (the bivariate distribution of the observed and missing data has been summed over the distribution of the missing data).

It is assumed that given a value of the latent variable the item responses are independent. Thus, the relationship among the item responses is accounted for by the latent variable. Given this assumption $f(\mathbf{y} \mid q_k, \mathbf{\Delta})$ can be written as

$$f(\mathbf{y} \mid q_k, \mathbf{\Delta}) = \prod_{j=1}^{J} P(q_k \mid \boldsymbol{\delta}_j)^{y_j}[1 - P(q_k \mid \boldsymbol{\delta}_j)]^{1-y_j} \,, \tag{2}$$

where $P(q_k \mid \boldsymbol{\delta}_j)$ is item characteristic curve for item j (which gives the probability of a correct response to item j as a function of the latent variable), and $\boldsymbol{\delta}_j$ are the item parameters for item j. From Equations 1 and 2 the likelihood of the observed data for a sample of N examinees is given by

$$L(\mathbf{Y} \mid \mathbf{\Delta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} P(q_k \mid \boldsymbol{\delta}_j)^{y_{ij}}[1 - P(q_k \mid \boldsymbol{\delta}_j)]^{1-y_{ij}} \right) \tag{3}$$

Maximum likelihood estimates of the parameters $\mathbf{\Delta}$ and $\boldsymbol{\pi}$ are the values that maximize Equation 3. Finding the parameters that maximize Equation 3 is sometimes called marginal maximum likelihood because the likelihood in Equation 3 is based on the marginal distribution of the observed data given in Equation 1. Finding maximum likelihood estimates using the observed data likelihood directly (Equation 3) can be complicated (Bock and Lieberman, 1970; Thissen, 1982). The EM algorithm greatly simplifies the computation of maximum likelihood estimates of $\mathbf{\Delta}$ and $\boldsymbol{\pi}$.

EM Algorithm

The EM algorithm is a method of finding the parameters that maximum the observed data likelihood given by Equation 3 using the *complete* data likelihood. The first step in describing the EM algorithm is to present the complete data likelihood which is used in the M step.

Complete Data Likelihood

The probability of a randomly sampled examinee having observed item responses $\mathbf{y}$ and a value of the latent variable in category k is

$$f(\mathbf{y}, q_k \mid \mathbf{\Delta}, \boldsymbol{\pi}) = f(\mathbf{y} \mid q_k, \mathbf{\Delta})\pi_k \,. \tag{4}$$

From Equations 4 and 2 the likelihood of the complete data for a sample of N examinees is

$$\begin{aligned} L(\mathbf{Y}, \boldsymbol{\theta} \mid \mathbf{\Delta}, \boldsymbol{\pi}) &= \prod_{i=1}^{N} \prod_{j=1}^{J} P(\theta_i \mid \boldsymbol{\delta}_j)^{y_{ij}}[1 - P(\theta_i \mid \boldsymbol{\delta}_j)]^{1-y_{ij}} f(\theta_i \mid \boldsymbol{\pi}) \\ &= \prod_{j=1}^{J} \prod_{i=1}^{N} P(\theta_i \mid \boldsymbol{\delta}_j)^{y_{ij}}[1 - P(\theta_i \mid \boldsymbol{\delta}_j)]^{1-y_{ij}} f(\theta_i \mid \boldsymbol{\pi}) \\ &= \prod_{j=1}^{J} \prod_{k=1}^{K} P(q_k \mid \boldsymbol{\delta}_j)^{r_{jk}}[1 - P(q_k \mid \boldsymbol{\delta}_j)]^{n_k - r_{jk}} \pi_k^{n_k} \,, \end{aligned} \tag{5}$$

where $f(\theta_i \mid \boldsymbol{\pi}) = \pi_k$ if $\theta_i = q_k$, n_k is the number of the N examinees for whom the latent variable is contained in category k, and r_{jk} is the number of examines for whom the latent variable is contained in category k and who answer item j correctly.

Another way of deriving the complete data likelihood is to recognize that the complete data consist of variables that are all discrete. There are $2^J K$ possible combinations of item responses and values of the latent variable. The data can be represented by the counts of the number of examinees that have each of these possible combinations of observed and latent variables. These counts have a multinomial distribution. Let $m_{\mathbf{y}q_k}$ be the number of examinees with observed item responses $\mathbf{y}$ and latent variable value q_k. Then the counts $m_{\mathbf{y}q_k}$ have a multinomial distribution with parameters given by the probabilities of the observed item responses being equal to $\mathbf{y}$ and the latent variable being equal to q_k. These multinomial probabilities depend on the parameters of the IRT model ($\mathbf{\Delta}$ and $\boldsymbol{\pi}$). The multinomial distribution $f(m_{\mathbf{y}q_k} \mid \mathbf{\Delta}, \boldsymbol{\pi})$ can be written as

$$f(m_{\mathbf{y}q_k} \mid \mathbf{\Delta}, \boldsymbol{\pi}) = f(m_{\mathbf{y}} \mid q_k, \mathbf{\Delta}, n_k) f(n_k \mid \boldsymbol{\pi}) \,, \tag{6}$$

where $m_{\mathbf{y}}$ is the number of examinees with item response pattern $\mathbf{y}$, $f(m_{\mathbf{y}} \mid q, \mathbf{\Delta}, n_k)$ is the distribution of the number of examinees with item response pattern $m_{\mathbf{y}}$ given the latent variable is equal to q_k, and $f(n_k \mid \boldsymbol{\pi})$ is the distribution of the number of examinees with latent variable equal to q_k. Both $f(m_{\mathbf{y}} \mid n_k, \mathbf{\Delta})$ and $f(n_k \mid \boldsymbol{\pi})$ are multinomial distributions (Bishop, Feinberg, and Holland, 1975, page 445). The likelihood of the counts $n_1, n_2, \ldots, n_K$, ignoring terms in the multinomial likelihood that do not depend on the parameters, is

$$L(n_1, n_2, \ldots, n_K \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{n_k} \,. \tag{7}$$

Since the item responses are independent given the latent variable the distribution $f(m_{\mathbf{y}} \mid q_k, \mathbf{\Delta}, n_k)$ can be written as a product of J binomial distributions giving the probability of r_{jk} successes in n_k trials with binomial probabilities $P(q_k \mid \boldsymbol{\delta}_j)$. The likelihood of the counts $\mathbf{r}_k = r_{1k}, r_{2k}, \ldots, r_{Jk}$, ignoring terms in the binomial likelihoods that do not depend on the parameters, is

$$L(r_{1k}, r_{2k}, \ldots, r_{Jk} \mid \mathbf{\Delta}, n_1, n_2, \ldots, n_K) = \prod_{j=1}^{J} P(q_k \mid \boldsymbol{\delta}_j)^{r_{jk}} [1 - P(q_k \mid \boldsymbol{\delta}_j)]^{n_k - r_{jk}} \,. \tag{8}$$

The likelihood of the counts $\mathbf{r}_k$ over all the latent variable categories is

$$L(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K \mid \mathbf{\Delta}, n_1, n_2, \ldots, n_K) = \prod_{k=1}^{K} \prod_{j=1}^{J} P(q_k \mid \boldsymbol{\delta}_j)^{r_{jk}} [1 - P(q_k \mid \boldsymbol{\delta}_j)]^{n_k - r_{jk}} \,. \tag{9}$$

The product of Equations 7 and 9 gives the likelihood of the counts $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K$ and $n_1, n_2, \ldots, n_K$ (ignoring some terms that do not depend on the parameters):

$$L(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K, n_1, n_2, \ldots, n_K \mid \mathbf{\Delta}, \boldsymbol{\pi}) = \prod_{j=1}^{J} \prod_{k=1}^{K} P(q_k \mid \boldsymbol{\delta}_j)^{r_{jk}} [1 - P(q_k \mid \boldsymbol{\delta}_j)]^{n_k - r_{jk}} \pi_k^{n_k} \,. \tag{10}$$

Equation 10 is the same as Equation 5. The complete data likelihood obtained from the counts $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K$ and $n_1, n_2, \ldots, n_K$ is the same as the likelihood obtained from the complete data for the individual examinees. The counts $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K$ and $n_1, n_2, \ldots, n_K$ are the complete data sufficient statistics for the parameters $\mathbf{\Delta}$ and $\boldsymbol{\pi}$.

It is more convenient to maximize the log of the likelihood rather than the likelihood. The log-likelihood (logarithm of Equation 10) is

$$\log[L(\mathbf{R}, \mathbf{n} \mid \mathbf{\Delta}, \boldsymbol{\pi})] = \sum_{j=1}^{J} \sum_{k=1}^{K} r_{jk} \log[P(q_k \mid \boldsymbol{\delta}_j)] + (n_k - r_{jk}) \log[1 - P(q_k \mid \boldsymbol{\delta}_j)] + n_k \log[\pi_k] \,, \tag{11}$$

where $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K)$ and $\mathbf{n} = (n_1, n_2, \ldots, n_K)$.

E Step and M Step

The EM algorithm is an iterative algorithm for estimating $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$ where there are two steps performed at each iteration: the E step and the M step.

E step. The E step at iteration s ($s = 0, 1, \ldots$) consists of computing the expected values of the complete data sufficient statistics $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K, n_1, n_2, \ldots, n_K)$ over the conditional distribution of the missing data given the observed data ($\mathbf{Y}$) and fixed values of the parameters $\boldsymbol{\Delta}^{(s)}$ and $\boldsymbol{\pi}^{(s)}$ obtained the M step of iteration $s-1$ (starting values for the parameters are used for $\boldsymbol{\Delta}^{(0)}$ and $\boldsymbol{\pi}^{(0)}$). The expected values of the complete data sufficient statistics computed at iteration s are denoted $r_{jk}^{(s)}, j = 1, \ldots J, k = 1, \ldots, K$ and $n_k^{(s)}, k = 1, \ldots, K$.

The conditional probability of the latent variable being equal to q_k for examinee i given observed item responses $\mathbf{y}_i$ and parameter values of $\boldsymbol{\Delta}^{(0)}$ and $\boldsymbol{\pi}^{(0)}$ is obtained from Equation 15 of Woodruff and Hanson (1996) and Equation 2:

$$\begin{aligned} f(q_k \mid \mathbf{y}_i, \boldsymbol{\Delta}^{(s)}, \boldsymbol{\pi}^{(s)}) &= \frac{f(\mathbf{y}_i \mid q_k, \boldsymbol{\Delta}^{(s)})\pi_k^{(s)}}{\sum_{k'=1}^K f(\mathbf{y}_i \mid q_{k'}, \boldsymbol{\Delta}^{(s)})\pi_{k'}^{(s)}} \\ &= \frac{\pi_k^{(s)} \prod_{j=1}^J P(q_k \mid \boldsymbol{\delta}_j^{(s)})^{y_{ij}}[1 - P(q_k \mid \boldsymbol{\delta}_j^{(s)})]^{1-y_{ij}}}{\sum_{k'=1}^K \pi_{k'}^{(s)} \prod_{j=1}^J P(q_{k'} \mid \boldsymbol{\delta}_j^{(s)})^{y_{ij}}[1 - P(q_{k'} \mid \boldsymbol{\delta}_j^{(s)})]^{1-y_{ij}}}\,, \end{aligned} \tag{12}$$

The value $n_k^{(s)}$ is the expected value of n_k over the conditional distribution given in Equation 12. This expected value is equal to the sum of the conditional probabilities of the latent variable for each examinee being equal to q_k. Using Equation 12 the value $n_k^{(s)}$ is given by

$$\begin{aligned} n_k^{(s)} &= E(n_k \mid \mathbf{Y}, \boldsymbol{\Delta}^{(s)}, \boldsymbol{\pi}^{(s)}) = \sum_{i=1}^N f(q_k \mid \mathbf{y}_i, \boldsymbol{\Delta}^{(s)}, \boldsymbol{\pi}^{(s)}) = \sum_{i=1}^N \frac{f(\mathbf{y}_i \mid q_k, \boldsymbol{\Delta}^{(s)})\pi_k^{(s)}}{\sum_{k'=1}^K f(\mathbf{y}_i \mid q_{k'}, \boldsymbol{\Delta}^{(s)})\pi_{k'}^{(s)}} \\ &= \sum_{i=1}^N \frac{\pi_k^{(s)} \prod_{j=1}^J P(q_k \mid \boldsymbol{\delta}_j^{(s)})^{y_{ij}}[1 - P(q_k \mid \boldsymbol{\delta}_j^{(s)})]^{1-y_{ij}}}{\sum_{k'=1}^K \pi_{k'}^{(s)} \prod_{j=1}^J P(q_{k'} \mid \boldsymbol{\delta}_j^{(s)})^{y_{ij}}[1 - P(q_{k'} \mid \boldsymbol{\delta}_j^{(s)})]^{1-y_{ij}}}\,. \end{aligned} \tag{13}$$

The value $r_{jk}^{(s)}$ is the expected value of r_{jk} over the conditional distribution given in Equation 12. This expected value is equal to the sum of the conditional probabilities of the latent variable for each examinee being equal to q_k for those examinees who answered item j correctly. From Equation 12 the value of $r_{jk}^{(s)}$ is given by

$$\begin{aligned} r_{jk}^{(s)} &= E(r_{jk} \mid \mathbf{Y}, \boldsymbol{\Delta}^{(s)}, \boldsymbol{\pi}^{(s)}) = \sum_{i=1}^N y_{ij} f(q_k \mid \mathbf{y}_i, \boldsymbol{\Delta}^{(s)}, \boldsymbol{\pi}^{(s)}) = \sum_{i=1}^N \frac{y_{ij} f(\mathbf{y}_i \mid q_k, \boldsymbol{\Delta}^{(s)})\pi_k^{(s)}}{\sum_{k'=1}^K f(\mathbf{y}_i \mid q_{k'}, \boldsymbol{\Delta}^{(s)})\pi_{k'}^{(s)}} \\ &= \sum_{i=1}^N \frac{y_{ij}\pi_k^{(s)} \prod_{j=1}^J P(q_k \mid \boldsymbol{\delta}_j^{(s)})^{y_{ij}}[1 - P(q_k \mid \boldsymbol{\delta}_j^{(s)})]^{1-y_{ij}}}{\sum_{k'=1}^K \pi_{k'}^{(s)} \prod_{j=1}^J P(q_{k'} \mid \boldsymbol{\delta}_j^{(s)})^{y_{ij}}[1 - P(q_{k'} \mid \boldsymbol{\delta}_j^{(s)})]^{1-y_{ij}}}\,. \end{aligned} \tag{14}$$

M step. The M step at iteration s consists of substituting the expected values of the complete data sufficient statistics obtained in the E step at iteration s in the complete data log-likelihood given by Equation 11 and finding the values of $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$ that maximize this log-likelihood. The values of $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$ that maximize the log-likelihood at iteration s are denoted $\boldsymbol{\Delta}^{(s+1)}$ and $\boldsymbol{\pi}^{(s+1)}$. These parameter estimates are used in the E step in iteration $s+1$.

The log-likelihood in Equation 11 using $r_{jk}^{(s)}$ and $n_j^{(s)}$ from iteration s of the E step can be written as

$$\log[L(\mathbf{R}^{(s)}, \mathbf{n}^{(s)} \mid \boldsymbol{\Delta}, \boldsymbol{\pi})] = \sum_{j=1}^J l(\boldsymbol{\delta}_j) + l(\boldsymbol{\pi})\,, \tag{15}$$

where

$$l(\boldsymbol{\delta}_j) = \sum_{k=1}^{K} r_{jk}^{(s)} \log[P(q_k \mid \boldsymbol{\delta}_j)] + (n_k^{(s)} - r_{jk}^{(s)}) \log[1 - P(q_k \mid \boldsymbol{\delta}_j)]\,, \tag{16}$$

and

$$l(\boldsymbol{\pi}) = \sum_{k=1}^{K} n_k^{(s)} \log[\pi_k]\,. \tag{17}$$

To find the value of a parameter that maximizes Equation 11 the derivative of Equation 11 with respect to the parameter is set equal to zero and solved for the parameter. The derivative of the complete data log-likelihood with respect to parameter t of item j (δ_{tj}) only depends on $l(\boldsymbol{\delta}_j)$ (if there are no common parameters across items), and the derivative of the complete data log-likelihood with respect to π_k only depends on $l(\boldsymbol{\pi})$. Consequently, maximum likelihood estimates of the parameters of each item and the parameters of the latent variable distribution can be computed separately. This is in contrast to computing maximum likelihood estimates using the observed data likelihood in Equation 3 where the derivative of the likelihood with respect to one parameter would involve all other parameters. When using the observed data likelihood all parameters must be estimated simultaneously. This is an example of the tremendous simplification in the computation of maximum likelihood estimates that can occur when using the EM algorithm. If some parameters are common across items then the parameters for items that have some common parameters cannot be separately estimated, but it is still the case that estimates of item parameters and the parameters of the latent variable distribution can be separately computed.

The portion of the log-likelihood corresponding to $\boldsymbol{\pi}$ $[l(\boldsymbol{\pi})]$ is the log-likelihood for a sample from a multinomial distribution with parameters $\boldsymbol{\pi}$. The maximum likelihood estimate of the multinomial probability π_k is n_k/N. Consequently, at iteration s in the M step the values of $\pi_k^{(s+1)}$ are computed by

$$\pi_k^{(s+1)} = \frac{n_k^{(s)}}{N}\,. \tag{18}$$

The M step calculations to obtain $\boldsymbol{\pi}^{(s+1)}$ are extremely simple.

The values of $\boldsymbol{\delta}_j^{(s+1)}$ computed in the M step at iteration s are the solution of the system of equations:

$$\frac{\partial l(\boldsymbol{\delta}_j)}{\partial \delta_{tj}} = 0\,, \tag{19}$$

for $t = 1, 2, \ldots, T_j$ where there are T_j parameters for item j. Substituting Equation 16 into Equation 19 and simplifying gives

$$\sum_{k=1}^{K} \frac{r_{jk}^{(s)} - n_k^{(s)} P(q_k \mid \boldsymbol{\delta}_j)}{[1 - P(q_k \mid \boldsymbol{\delta}_j)] P(q_k \mid \boldsymbol{\delta}_j)} \frac{\partial P(q_k \mid \boldsymbol{\delta}_j)}{\partial \delta_{tj}} = 0\,, \tag{20}$$

for $t = 1, 2, \ldots, T_j$. Solving the system of equations in Equation 20 for δ_{tj}, $t = 1, 2, \ldots, T_j$ will result in $\boldsymbol{\delta}^{(s+1)}$. Iterative procedures such as Newton-Raphson (Dennis and Schnabel, 1983) will typically be needed to solve the system of equations given by Equation 20.

Implementation of the EM Algorithm for IRT Models

This section describes the specific procedure used to apply the EM algorithm in computing maximum likelihood estimates of $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$ in IRT models using the results presented above. At iteration s, $s = 0, 1, \ldots$, the procedures used for the E step and M step are:

E step. Substitute $\boldsymbol{\pi}^{(s)}$ and $\boldsymbol{\Delta}^{(s)}$ computed in iteration $s - 1$ (or starting values in iteration 0) in Equations 13 and 14 to produce values of $n_k^{(s)}$ and $r_{jk}^{(s)}$.

M step. The M step is performed in two parts which separately produce values of $\boldsymbol{\pi}^{(s+1)}$ and $\boldsymbol{\Delta}^{(s+1)}$:

1. Compute values of $\pi_k^{(s+1)}$, $k = 1, \ldots, K$, using Equation 18 and the values of $n_k^{(s)}$ computed in the E step.

2. Compute values of $\boldsymbol{\delta}_j^{(s+1)}$, $j = 1, 2, \ldots J$, by solving the system of equations for each j given by Equation 20 using the values of $r_{jk}^{(s)}$ and $n_k^{(s)}$ computed in the E step.

Iterations of the E and M steps are repeated until the parameter estimates convergence. Convergence can be assessed by relative difference in the observed data likelihood from one iteration to the next (the EM algorithm guarantees the observed data likelihood will increase on each iteration), or by differences in parameter estimates between iterations.

The only part of the EM algorithm that could differ significantly for different IRT models is part 2 of the M step. The degree of difficulty in solving the system of equations in part 2 of the M step for the item parameters can vary for different IRT models.

The description of part 2 of the M step assumes there are no common item parameters across items. If there are common item parameters across items (e.g., an item parameter is set to be equal across two or more items) then the derivative of $\sum_{j=1}^{J} \boldsymbol{\delta}_j$ in the log-likelihood of Equation 15 will depend on more than just one $\boldsymbol{\delta}_j$, and the system of equations to be solved for the maximum likelihood estimates of the item parameters will be more complicated than that given in Equation 20.

Note that the EM algorithm can be used to just estimate the latent variable distribution for a fixed set of item parameters by only carrying out the first part of the M step on each iteration. Similarly, just the item parameters can be estimated for a fixed latent variable distribution by only carrying out the second part of the M step in each iteration.

Example — Guttman Scale Model

This section describes details of using the EM algorithm to find maximum likelihood parameter estimates for a specific IRT model. The model to be considered is a generalized Guttman scale model, which is a type of latent class model. For a latent class model the discrete levels of the latent variable $(q_1, q_2, \ldots, q_K)$ are assumed to be nominal categories. The values of q_k are just labels rather than numerical values that represent a position of an examinee on an underlying latent numerical scale. For the Guttman scale model the latent classes will be referred to as levels (implying an ordering of the classes). The K levels (latent classes) are labeled $0, 1, \ldots, K-1$ (so $q_k = k-1$). The levels are ordered in the following sense. Examinees at level 0 cannot answer any of the items correctly. Each level $q_k, q_k > 0$ has m_k items associated with it. Examinees at level q_k can correctly answer all items at level q_k and lower, but cannot correctly answer items at level q_{k+1} and higher. In other words, examinees at level q_k have all the skills necessary to correctly answer items at level q_k and lower, but additional skills not possessed by examinees at level q_k are needed to correctly answer items at level q_{k+1} and higher. This describes a generalized Guttman scale model where more than one item is allowed at each level. In a traditional Guttman scale model there is only one item per level.

Associated with each level is a latent response pattern which indicates for each item whether examinees at that level can answer the item correctly or not. For example, at level 0 the latent response pattern would be a vector of zeros indicating that examinees at level 0 cannot answer any of the items correctly. The latent response pattern at level 2 would have ones for items at levels 1 and 2, and zeros for items at levels 3 and above. If there is only one item associated with level $q_k, k = 1, 2, \ldots, K$, then the latent response patterns form a traditional Guttman scale.

The latent response patterns indicate the responses of examinees at each level if examinees made no errors. Response patterns other than the latent response patterns associated with each level can occur if examinees make errors in responding to items. There are two types of errors that can occur — 1) an examinee who should answer an item correctly answers it incorrectly (false negative error), and 2) an examinee who should not be able to answer an item correctly answers it correctly (false positive error). To account for these types of errors each item has two error probabilities associated with it. The false negative error rate associated with item j is denoted α_j, and the false positive error rate associated with item j is denoted β_j. Consequently, the probability that an examinee at level q_k correctly answers item j $[P(q_k \mid \alpha_j, \beta_j)]$ is

$$P(q_k \mid \alpha_j, \beta_j) = v_{kj}(1 - \alpha_j) + (1 - v_{kj})\beta_j \,, \tag{21}$$

where $\mathbf{v}_k = (v_{k1}, v_{k2}, \ldots, v_{kJ})$ is the latent response pattern for level q_k. The probability that an examinee at level q_k answers item j incorrectly is

$$1 - P(q_k \mid \alpha_j, \beta_j) = v_{kj}\alpha_j + (1 - v_{kj})(1 - \beta_j) \,. \tag{22}$$

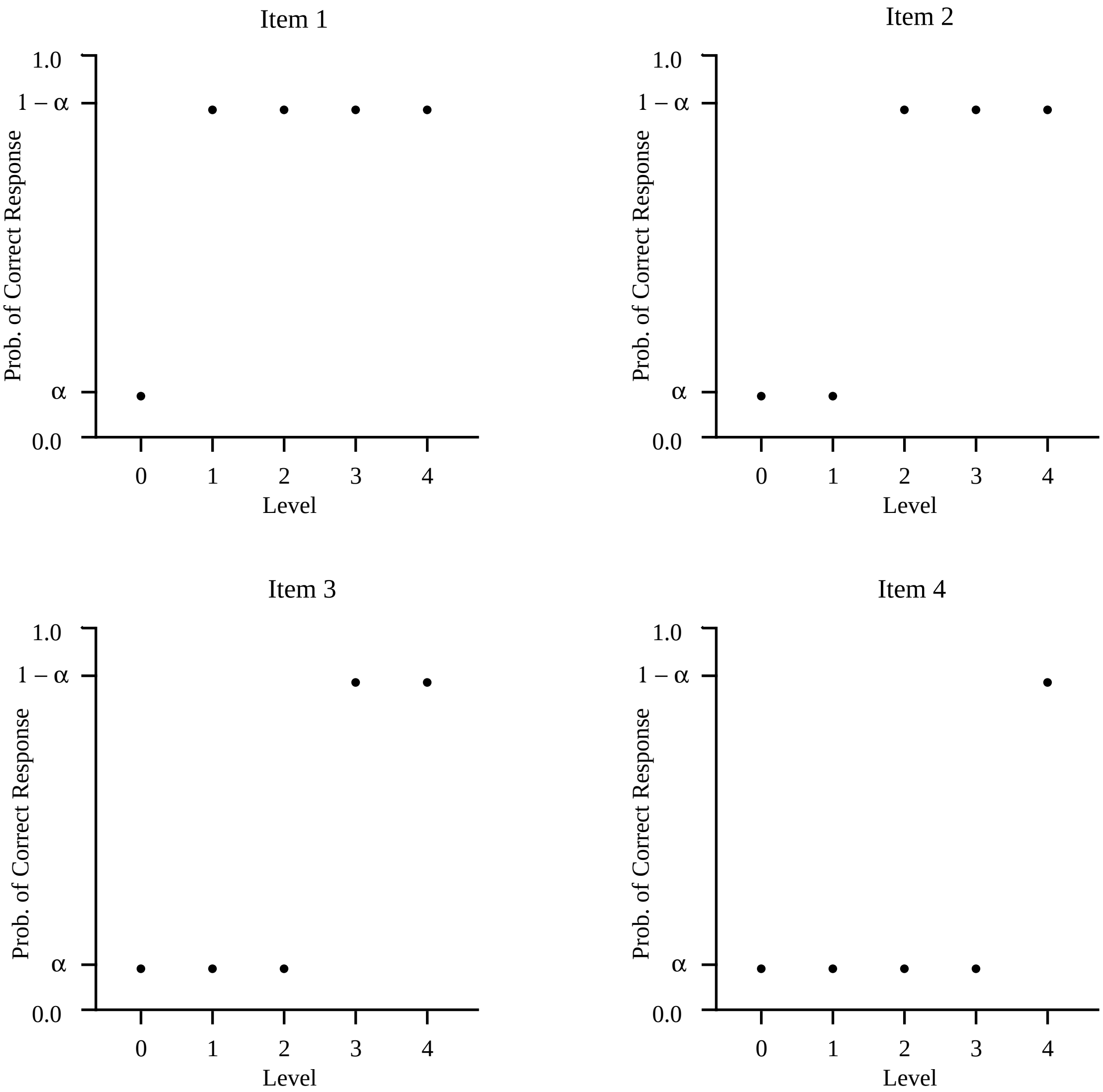

Figure 1. Item Response Probabilities for Proctor's Constant Error Rate Model.

Guttman scale models with restrictions on the false negative and false positive error rates given in Equations 21 and 22 can be used. One type of restriction is that $\alpha_j = \beta_j$. In this case the false negative and false positive error rates are constant for item j (the error rates are constant within an item). Another type of restriction is that $\alpha_j = \alpha$ and $\beta_j = \beta$. In this case the false negative and false positive error rates are constant between items. The two types of restrictions can be combined. If error rates are constant both within and between items this means the false positive and false negative error rates for all items are equal to a constant value ($\alpha_j = \beta_j = \alpha$, for all j).

Proctor (1970) presented a Guttman scale model for the case of one item per level (a traditional Guttman scale) that had constant error rates both between and within items. Figure 1 presents plots of the probability of correct response for a four item test assuming Proctor's constant error rate model (where item j is associated with level j, $j = 1, \ldots 4$). The plots in Figure 1 are analogous to item characteristic curves in latent trait models where the probability of a correct response to an item is a function of a continuous latent variable. In Figure 1 the probability of a correct response for each item is a function of the ordered discrete levels.

Maximum Likelihood Estimates using the EM Algorithm

Everitt (1984) and Bartholomew (1987) discuss using the EM algorithm to estimate parameters of latent class models. The EM algorithm as described above can be used to compute maximum likelihood estimates for the Guttman scale model. The only thing that needs to be specified are the systems of equations solved for the parameters of each item in the M step given by Equation 20. Once these systems of equations are specified the EM algorithm can be carried out as described.

For item j there exists a k_j^* such that $v_{kj} = 0$ for $q_k < q_{k_j^*}$ and $v_{kj} = 1$ for $q_k \geq q_{k_j^*}$ (in the case of one item per level, where item j is at level j, $k_j^* = j$). Therefore, from Equation 21 $P(q_k \mid \alpha_j, \beta_j) = 1 - \alpha_j$ for $q_k < q_{k_j^*}$ and $P(q_k \mid \alpha_j, \beta_j) = \beta_j$ for $q_k \geq q_{k_j^*}$. Thus, Equation 20 with the derivative taken with respect to parameter α_j can be written as

$$\sum_{k=1}^{k_j^*-1} \frac{r_{jk}^{(s)} - n_k^{(s)}\beta_j}{(1-\beta_j)\beta_j} \frac{\partial \beta_j}{\partial \alpha_j} + \sum_{k=k_j^*}^{K} \frac{r_{jk}^{(s)} - n_k^{(s)}(1-\alpha_j)}{(1-\alpha_j)\alpha_j} \frac{\partial(1-\alpha_j)}{\partial \alpha_j} = 0\,. \tag{23}$$

Since $\partial\beta_j/\partial\alpha_j = 0$ and $\partial(1-\alpha_j)/\partial\alpha_j = -1$ Equation 23 can be written as

$$\sum_{k=k_j^*}^{K} \frac{-r_{jk}^{(s)} + n_k^{(s)}(1-\alpha_j)}{[1-\alpha_j]\alpha_j} = \frac{1}{(1-\alpha_j)\alpha_j}\left((1-\alpha_j)\sum_{k=k_j^*}^{K} n_k^{(s)} - \sum_{k=k_j^*}^{K} r_{jk}^{(s)}\right) = 0\,. \tag{24}$$

Solving Equation 24 for α_j gives

$$\begin{aligned}(1-\alpha_j)\sum_{k=k_j^*}^{K} n_k^{(s)} &= \sum_{k=k_j^*}^{K} r_{jk}^{(s)} \\ 1-\alpha_j &= \frac{\sum_{k=k_j^*}^{K} r_{jk}^{(s)}}{\sum_{k=k_j^*}^{K} n_k^{(s)}} \\ \alpha_j &= 1 - \frac{\sum_{k=k_j^*}^{K} r_{jk}^{(s)}}{\sum_{k=k_j^*}^{K} n_k^{(s)}}\,. \end{aligned} \tag{25}$$

Equation 20 with the derivative taken with respect to parameter β_j can be written as

$$\sum_{k=1}^{k_j^*-1} \frac{r_{jk}^{(s)} - n_k^{(s)}\beta_j}{(1-\beta_j)\beta_j} \frac{\partial \beta_j}{\partial \beta_j} + \sum_{k=k_j^*}^{K} \frac{r_{jk}^{(s)} - n_k^{(s)}(1-\alpha_j)}{(1-\alpha_j)\alpha_j} \frac{\partial(1-\alpha_j)}{\partial \beta_j} = 0\,, \tag{26}$$

or

$$\sum_{k=1}^{k_j^*-1} \frac{r_{jk}^{(s)} - n_k^{(s)}\beta_j}{(1-\beta_j)\beta_j} = 0\,. \tag{27}$$

Solving Equation 27 for β_j gives

$$\begin{aligned}\beta_j \sum_{k=1}^{k_j^*-1} n_k^{(s)} &= \sum_{k=1}^{k_j^*-1} r_{jk}^{(s)} \\ \beta_j &= \frac{\sum_{k=1}^{k_j^*-1} r_{jk}^{(s)}}{\sum_{k=1}^{k_j^*-1} n_k^{(s)}}\,. \end{aligned} \tag{28}$$

For the Guttman scale model part 2 of the M step involves the simple computations given by Equations 25 and 28. In this case computing the item parameter estimates in part 2 of the M step is as simple as computing the latent variable distribution in part 1 of the M step. For a Guttman scale model the M step is extremely easy to implement, and each iteration of the EM algorithm involves a sequence of simple computations which are easy to program.

Bayes Modal Estimates and Polytomous Items

This paper has described the EM algorithm for computing maximum likelihood parameter estimates for dichotomous IRT models. This section gives brief presentations of two additional topics in the application of the EM algorithm to computing item parameters in IRT models: 1) using the EM algorithm to compute Bayes modal estimates, and 2) using the EM algorithm to compute maximum likelihood estimates for polytomous IRT models.

Bayes Modal Estimates

The EM algorithm can be used to compute Bayes modal estimates as well as maximum likelihood estimates (Dempster, Laird, and Rubin, 1977; Tanner, 1996). This section only discusses computing Bayes modal estimates of item parameters. Hanson (1998) discusses using the EM algorithm to compute Bayes modal estimates of the latent variable distribution.

The only difference in the EM algorithm described for computing maximum likelihood estimates of item parameters and the EM algorithm used to compute Bayes modal estimates occurs in part 2 of the M step. Instead of finding parameter estimates that maximize the complete data likelihood, parameter estimates are found that maximize the complete data posterior distribution. In this description it is assumed that the prior distributions for all item parameters are independent. In this case the logarithm of the complete data posterior analogous to the log-likelihood in Equation 15 is:

$$\log[L(\mathbf{R}^{(s)}, \mathbf{n}^{(s)} \mid \boldsymbol{\Delta}, \boldsymbol{\pi})] = \sum_{j=1}^{J} l(\boldsymbol{\delta}_j) + \sum_{j=1}^{J} \sum_{t=1}^{T_j} \log[g(\delta_{tj})] + l(\boldsymbol{\pi}) \,, \tag{29}$$

where $g(\delta_{tj})$ is the prior distribution of item parameter δ_{tj}, and $l(\boldsymbol{\delta}_j)$ and $l(\boldsymbol{\pi})$ are given by Equations 16 and 17, respectively. The system of equations analogous to Equation 20 to be solved in part 2 of the M step are

$$\frac{\partial \log[g(\delta_{tj})]}{\partial \delta_{tj}} + \sum_{k=1}^{K} \frac{r_{jk}^{(s)} - n_k^{(s)} P(q_k \mid \boldsymbol{\delta}_j)}{[1 - P(q_k \mid \boldsymbol{\delta}_j)] P(q_k \mid \boldsymbol{\delta}_j)} \frac{\partial P(q_k \mid \boldsymbol{\delta}_j)}{\partial \delta_{tj}} = 0 \,, \tag{30}$$

for $t = 1, 2, \ldots, T_j$.

The EM algorithm for Bayes modal estimates is the same as the EM algorithm described for maximum likelihood estimates except that the system of equation solved in part 2 of the M step are given by Equation 30 rather than Equation 20.

Polytomous Items

The EM algorithm described above for dichotomous items can be generalized to polytomous items. If there are L_j response categories $(0, 1, \ldots, L_j - 1)$ for item j then the conditional probability of a response in category l, $l = 0, 1, \ldots, L_j - 1$, of item j given a latent variable value of q_k is the item category response function $P_l(q_k \mid \boldsymbol{\delta}_j)$. The complete data log-likelihood analogous to Equation 11 in the case of polytomous items is (Woodruff and Hanson, 1996):

$$\sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{l=0}^{L_j - 1} \log[P_l(q_k, \boldsymbol{\delta}_j)] r_{jkl}^{(s)} + \sum_{k=1}^{K} n_k \log[\pi_k] \,, \tag{31}$$

where $r_{jkl}^{(s)}$ is the number of examinees who have a latent variable value of q_k and respond in category l of item j. The complete data sufficient statistics that are computed in the E step at iteration s are $r_{jkl}^{(s)}$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$, $l = 1, 2, \ldots, L_j$. It is shown in Woodruff and Hanson (1996) that

$$n_k^{(s)} = \sum_{l=0}^{L_j - 1} r_{jkl}^{(s)} \,, \tag{32}$$

for all j.

The EM algorithm in the case of polytomous items consists of computing $r_{jkl}^{(s)}$ in the E step and then finding the parameters that maximize the complete data log-likelihood in Equation 31 subject to the constraint that

$$\sum_{l=0}^{L_j-1} P_l(q_k, \boldsymbol{\delta}_j) = 1\,, \tag{33}$$

for $J = 1, 2, \ldots, J$ and $K = 1, 2, \ldots, K$. For more details on using the EM algorithm to compute parameter estimates for polytomous IRT models see Woodruff and Hanson (1996, 1997).

References and EM Bibliography

Baker, F. B. (1992). *Item response theory parameter estimation techniques.* New York: Marcel Dekker.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis.* New York: Oxford University Press.

Bishop, Y. M. M, Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate distributions.* Cambridge, MA: MIT Press.

Bock, R. D.,& Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179-197.

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

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

Dennis, J. E. & Schnabel, R. B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations.* Englewood Cliffs, N.J.: Prentice-Hall.

Everitt, B. S. (1984). A note on parameter estimation for Lazarsfeld's latent class model using the EM algorithm. *Multivariate Behavioral Research*, *19*, 79-89.

Hanson, B. A. (1998). *Bayes modal estimates of a discrete latent variable distribution in item response models using the EM algorithm.* Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, April). [Available at http://www.b-a-h.com/papers/paper9801.html]

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics*, *13*, 243-271.

Harwell, M. R., Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement*, *15*, 375-389.

Hsu, Y., & Fan, M. (1997). *The relationship between the Bock-Aitkin procedure and the EM algorithm for IRT model estimation.* Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, March).

McLachlan, G. J. & Krishnan, T. (1997). *The EM algorithm and extensions.* New York: John Wiley and Sons.

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

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis for Guttman scaling. *Psychometrika*, *35*, 73-78.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika*, *48*, 567-574.

Tanner, M. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (3rd ed.). New York: Springer-Verlag.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175-186.

Tsutakawa, R. K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics*, *9*, 263-276.

Tsutakawa, R. K., & Lin H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, *51*, 251-267.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of*

Educational Statistics, *13*, 117-130.

Wilson, M., & Adams, R. J. (1993). Marginal maximum likelihood estimation for the ordered partition model. *Journal of Educational Statistics*, *18*, 69-90.

Woodruff, D. J., & Hanson, B. A. (1996). *Estimation of item response models using the EM algorithm for finite mixtures.* ACT Research Report 96-6. Iowa City, IA: American College Testing.

Woodruff, D. & Hanson, B. A. (1997). *Estimation of item response models using the EM algorithm for finite mixtures.* Paper presented at the Annual Meeting of the Psychometric Society (Gatlinburg, Tennessee, June). [Available at http://www.b-a-h.com/papers/paper9701.html]