HUDM 6052: Psychometric Theory II

Lecture 8: Marginal Maximum Likelihood Estimation (MMLE)

(Baker & Kim (2004), Chapter 6, Embretson & Reise (2000), Chapter 8)

Recall that the limitation of JMLE is that the item parameter estimates are not consistent. This is because as the sample size increases, so do the number of unknown ability parameters θ. One way around this problem is CMLE. In JMLE we maximize the log-likelihood function conditional upon the unknown item and ability parameters. In CMLE we maximize the log-likelihood function conditional upon the unknown item parameters and the test scores of examinees. CMLE works for the Rasch model because the test scores are sufficient statistics for the unknown ability parameters. But for models like the 2PL or 3PL, there are no sufficient statistics that can be derived strictly from the data, so CMLE is limited to the Rasch model. The basic idea in marginal maximum likelihood estimation (MMLE) is to maximize a marginal log-likelihood function in which the unknown abilities have been integrated out. MMLE can be applied with all of the IRT models for dichotomously scored data.

The MMLE approach is based on a result using an application of Bayes' theorem:

$$P(\theta_j|U_j,\tau,\xi) = \frac{P(U_j|\theta_j,\xi)g(\theta|\tau)}{\int_\theta P(U_j|\theta_j,\xi)g(\theta|\tau)} \qquad (1)$$

In the numerator, the first term $P(U_j|\theta_j,\xi) = \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}}$ is the likelihood function of the U_j response pattern for an examinee of ability θ_j, while $g(\theta|\tau)$ represents the density of θ (having parameters τ) in the population of examinees. The denominator $\int_\theta P(U_j|\theta_j,\xi)g(\theta|\tau)$ is the marginal probability of the response pattern U_j computed over the density of θ.

Posterior distribution (1) combines the information from the prior θ distribution and that in the likelihood function. The prior distribution essentially does distribute the data across the θ scale in proportion to the posterior probability of examinees being at each point on the scale, i.e., it determines the expected number of examinees at each point along the θ scale.

<u>Bock & Lieberman's MML solution</u>

The likelihood and log-likelihood functions of the full data matrix with respect to the unknown item and ability parameters are:

$$L = \prod_{j=1}^{N} P(U_j) \qquad (2)$$

$$\log L = \prod_{j=1}^{N} \log P(U_j)$$

As with the other estimation procedures, we differentiate log L with respect to each unknown item parameter and use Newton-Raphson to find the maximum likelihood estimates. For example, in the 3PL model, we obtain the following for the derivative computed with respect to an item discrimination parameter a_i :

$$\frac{\partial}{\partial a_i}(\log L) = \sum_{j=1}^{N} \frac{\partial}{\partial a_i}(\log P(U_j)) = \sum_{j=1}^{N} [P(U_j)]^{-1} \int_{\theta} \frac{\partial}{\partial a_i}[P(U_j|\theta_j, \xi)] g(\theta|\tau) d\theta$$

It is known that

$$\frac{\partial}{\partial a_i}[P(U_j|\theta_j, \xi)] = \frac{\partial}{\partial a_i}[\log P(U_j|\theta_j, \xi)] P(U_j|\theta_j, \xi)$$

This produces:

$$\frac{\partial}{\partial a_i}(\log L) = \sum_{j=1}^{N} [P(U_j)]^{-1} \int_{\theta} \frac{\partial}{\partial a_i}[P(U_j|\theta_j, \xi)] P(U_j|\theta_j, \xi) g(\theta|\tau) d\theta$$

$$= \sum_{j=1}^{N} \int_{\theta} \frac{\partial}{\partial a_i}[\log P(U_j|\theta_j, \xi)] \left[\frac{P(U_j|\theta_j, \xi) g(\theta|\tau)}{P(U_j)} \right] d\theta$$

so that

$$\frac{\partial}{\partial a_i}(\log L) = \sum_{j=1}^{N} \int_{\theta} \frac{\partial}{\partial a_i}[\log P(U_j|\theta_j, \xi)][P(\theta_j|U_j, \tau, \xi)] d\theta$$

After working through the derivatives of $P(U_j)$ with respect to a_i, this derivative reduces to

$$\frac{\partial}{\partial a_i}(\log L) = (1 - c_i) \sum_{j=1}^{N} \int_{\theta} [u_{ij} - P_i(\theta_j)] W_{ij} (\theta_j - b_i) [P(\theta_j|U_j, \tau, \xi)] d\theta = 0 \qquad (3)$$

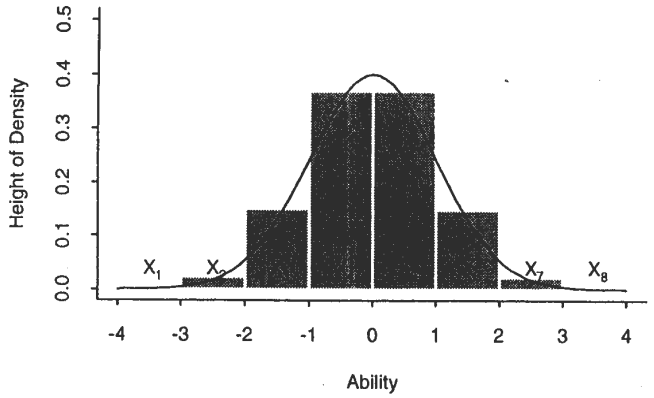
where $W_{ij} = \frac{P_i^*(\theta_j)Q_i^*(\theta_j)}{P_i(\theta_j)Q_i(\theta_j)}$, and P_i^* and Q_i^* are as before just the logistic part of the 3PL model.

Similar expressions can be derived for the difficulty b_i and guessing c_i parameters in the 3PL model. The unique aspect of this estimation approach relates to the integration over θ that occurs in evaluating each derivative. This is what we mean by the fact that the maximum likelihood procedure is a *marginal* maximum likelihood procedure.

But the problem with (3) is that the integral is difficult to evaluate, given that θ is continuous. What is needed to make the above expression useful is a way of approximating it. To do this, a quadrature approximation method is used.

Quadrature distributions

The problem of finding the area under a continuous curve is replaced by the simpler problem of finding the sum of a finite number of rectangles that approximate the area under the curve.



$_k$ (k = 1, 2, …, q) is called a "node."
; has an associated weight $A(X_k)$ that
$(\theta|\tau)$ in the neighborhood of X_k and the

ɜ computer software, the derivative in (3)

$$\frac{\partial}{\partial a_i}(\log L) = (1-c_i)\sum_{k=1}^{} \sum_{j=1} [u_{ij} - r_i(X_k)]w_{ik}(X_k - b_i)[P(X_k|U_j,\tau,\xi)] = 0$$

Similar expressions result for the b_i and c_i parameters (refer to pp. 179-180 in Baker & Kim).

Disadvantages of the Bock & Lieberman approach:

- The Newton-Raphson iterations work with all items at the same time, therefore requiring the inversion of a very large matrix when there are a large number of items. This limits the Bock & Lieberman approach to cases involving a small number of items.

- Assumes the distribution of θ is known in advance.

Bock & Aitkin solution

The Bock & Aitkin method differs in two primary ways from the Bock & Lieberman approach:

- Like the Birnbaum's JMLE approach, it assumes items and examinees are mutually independent so that the Newton-Raphson procedure can be performed one item at a time.
- It estimates the ability distribution at the same time that it estimates item parameters.

The equivalent of (2) evaluated with respect to X_k rather than θ is

$$L(X_k) = \prod_{i=1}^{n} P_i(X_k)^{u_{ij}} Q_i(X_k)^{1-u_{ij}} \qquad (4)$$

and the equivalent of (1) becomes

$$P(X_k | U_j, \tau, \xi) = \frac{L(X_k)A(X_k)}{\sum_k^q L(X_k)A(X_k)} \qquad (5)$$

.

Bock and Aitkin's solution is implemented using an EM algorithm. The EM algorithm is an iterative procedure for finding maximum likelihood estimates of parameters of probability model in the presence of unobserved random variables. In IRT, we want maximum likelihood estimates of the item parameters for an item response model that has unobservable θ s. To make inference about θ, some observable representation based on the data matrix U is used. We think of (U,θ) as the unobserved (complete) data, while (U) is the observed (incomplete) data.

Let $f(U,\theta|\xi)$ represent the joint probability density function of the complete data where ξ consists of the item parameters to be estimated. Given the matrix of provisional item parameters at the p th cycle of the algorithm, ξ^{p+1} is determined by maximizing the posterior expectation $(E[\log f(U,\theta|\xi)|U,\xi^p])$ with respect to ξ. This process is repeated until a convergence criterion is satisfied.

The two steps of the EM approach can be summarized as:

- E-step – Compute $E[\log f(U,\theta|\xi)|U,\xi^p]$.
- M-step – Choose ξ^{p+1} such that the posterior expectation is maximized.

The EM strategy as implemented for the Bock-Aitkin solution involves computation of two important quantities:

$$\bar{f}_{ik} = \sum_{j}^{N}\left[\frac{L(X_k)A(X_k)}{\sum_{k}^{q}L(X_k)A(X_k)}\right] \qquad (6)$$

$$\bar{r}_{ik} = \sum_{j}^{N}\left[\frac{u_{ij}L(X_k)A(X_k)}{\sum_{k}^{q}L(X_k)A(X_k)}\right] \qquad (7)$$

In quadrature terms, these correspond roughly to the terms $g(\theta_j|\tau)$ and $P_i(\theta_j)$ given earlier. $\bar{f}_{ik}$ is the number of examinees at quadrature point X_k for item i, while $\bar{r}_{ik}$ is the number of examinees at quadrature point X_k that answer item i correctly.

- E-step
  - Use the quadrature form in (4) and provisional item parameter estimates to compute the likelihood of each examinee's item score vector at each of the q quadrature nodes.
  - Use expression (5) and the quadrature weights at each of the q nodes to compute the posterior probability that the ability of the j th examinee is X_k.
  - Use equations (6) and (7) to generate $\bar{f}_{ik}$ and $\bar{r}_{ik}$, the expected numbers of examinees attempting item i and the number of correct responses for that item at each of the q ability nodes.
- M-step

  Solve equations below for the item parameter estimates using $\bar{f}_{ik}$ and $\bar{r}_{ik}$ as the "artificial data." (and a Taylor series and Newton-Raphson procedure). The marginal likelihood equations (the equivalent equation for the a parameter) can be written as follows:

$$\frac{\partial}{\partial a_i}(\log L) = (1-c_i)\sum_{k=1}^{q}(X_k - b_i)[\bar{r}_{ik} - \bar{f}_{ik}P_i(X_k)]W_{ik} = 0$$

$$\frac{\partial}{\partial b_i}(\log L) = (-a_i)(1-c_i)\sum_{k=1}^{q}[\bar{r}_{ik} - \bar{f}_{ik}P_i(X_k)]W_{ik} = 0$$

$$\frac{\partial}{\partial c_i}(\log L) = (1-c_i)^{-1}\sum_{k=1}^{q}\frac{([\bar{r}_{ik} - \bar{f}_{ik}P_i(X_k)]}{P_i(X_k)} = 0$$

- Repeat the process until the likelihood no longer changes much from iteration to iteration.

BILOG-MG program (Zimowski, Muraki, Mislevy, & Bock, 1996)

```
EXAMPLE BILOG RUN: THREE PARAMETER MODEL,
     SOCIAL DESIRABILITY DATA
>COMMENTS
>GLOBAL    NPARM=3, DFName='C:\Program Files\bilogmg\SOCDES.DAT';
>LENGTH    NITems=20;
>INPUT     NTOtal=20, NALt=5, NIDchar=10;
(4X, 10A1, T1, 20A1)
>TEST      TNAme=SCODES;
>CALIB     NQP=20, CYCLES=10, NEWTON=2, CRIT=.01, IDI=0;
>SCORE     RSCTYPE=3;
```