Lecture 6: Joint Maximum Likelihood Estimation

(Baker & Kim (2004), Chapter 4; Embretson & Reise (2000), Chapter 8)

We consider three procedures for estimating item parameters in the presence of unknown abilities: *joint maximum likelihood estimation* (JMLE)*, conditional maximum likelihood estimation* (CMLE)*,* and *marginal maximum likelihood estimation* (MMLE). In all cases the likelihood and log-likelihood functions are defined as:

$$\text{Prob}(U|\theta) = \prod_{j=1}^{N}\prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}} = \prod_{j=1}^{N}\prod_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}$$

$$L = \log \text{Prob}(U|\theta) = \sum_{j=1}^{N}\sum_{i=1}^{n} [u_{ij} \log P_{ij} + (1-u_{ij}) \log Q_{ij}]$$

Compared to the log-likelihood functions considered before, notice L is now evaluated across the entire item response dataset: all N examinees and all n items. Assuming we want to fit a 2PL model, we want to determine $\hat{\zeta}$ and $\hat{\lambda}$ for each item, implying a total of $2n$ item parameters that need to be estimated: $\hat{\zeta}_1$, $\hat{\lambda}_1$; $\hat{\zeta}_2$, $\hat{\lambda}_2$; …; $\hat{\zeta}_n$, $\hat{\lambda}_n$. At the same time, there are a total of N ability parameters to be considered: $\hat{\theta}_1$, $\hat{\theta}_2$, ..., $\hat{\theta}_N$.

As result, application of the Newton-Raphson approach becomes much more complicated. We have $2n+N$ parameters, and since we need to differentiate L with respect to each and find an estimate at which the derivative is 0, we have a total of $2n+N$ equations to solve. That is, if we want to maximize the log-likelihood with respect to all of the model parameters, we need to compute $\frac{\partial L}{\partial \zeta_1}$, $\frac{\partial L}{\partial \lambda_1}$, …, $\frac{\partial L}{\partial \zeta_n}$, $\frac{\partial L}{\partial \lambda_n}$, and $\frac{\partial L}{\partial \theta_1}$,...,$\frac{\partial L}{\partial \theta_N}$. In addition, we need to compute the second partial derivatives with respect to all pairs of parameters, which are of the form: $\frac{\partial^2 L}{\partial \zeta_i \zeta_j}$, $\frac{\partial^2 L}{\partial \lambda_i \lambda_j}$, $\frac{\partial^2 L}{\partial \zeta_i \lambda_j}$, $\frac{\partial^2 L}{\partial \zeta_i \theta_k}$, $\frac{\partial^2 L}{\partial \lambda_i \theta_k}$, and $\frac{\partial^2 L}{\partial \theta_k \theta_l}$.

The Newton-Raphson equation takes the form

$$A_{t+1} = A_t - B_t^{-1} \cdot F_t$$

where

- A = column vector of item and ability parameter estimates of length $2n+N$
- B = matrix of second-order partial derivatives of dimension $2n+N \times 2n+N$

- F = column vector of the first-order derivatives of the likelihood function of length $2n + N$
- t = index of the iteration

Because the dimensions of B are so large, unless some assumptions are made concerning its elements, this procedure is too computationally demanding. So the following assumptions are made, all of which are likely to be reasonable for most item response data sets:

- $\frac{\partial^2 L}{\partial \theta_k \theta_l} = 0$ whenever $k \neq l$. Because the examinees can usually be regarded as an independent sample from a population of respondents, their ability estimates should be independent.

- $\frac{\partial^2 L}{\partial \zeta_i \theta_k} = \frac{\partial^2 L}{\partial \lambda_i \theta_k} = 0$. This implies that examinees and items are independent.

- $\frac{\partial^2 L}{\partial \zeta_i \zeta_j} = \frac{\partial^2 L}{\partial \lambda_i \lambda_j} = \frac{\partial^2 L}{\partial \zeta_i \lambda_j} = 0$ whenever $i \neq j$. This implies that the parameters for all items are independent.

These three assumptions eliminate most of the elements in the information matrix and simplify computations considerably. But the matrix is still of dimension $2n + N$. A procedure proposed by Birnbaum takes advantage of the assumption that items and examinees are all mutually independent and uses a two-stage estimation procedure:

- Step 1: Assume ability estimates are known. Then for each item, execute one step of the Newton-Raphson strategy for estimating the parameters of an item assuming known ability parameters.

- Step 2: Using the updated item parameter estimates from the previous step, assume they are the known item parameters. Then for each examinee, execute one step of the Newton-Raphson strategy to find a better estimate of ability.

- Step 3: Assume the ability estimates from the previous step (Step 2) are the known abilities. Because the metric of the ability scale is only unique up to a linear transformation, the abilities and item parameters are usually rescaled so that the ability distribution has a mean of 0 and a variance of 1. There are a number of different ways of doing this, but the following transformation is preferred:

$$\hat{\theta}_j^* = \frac{(\hat{\theta}_j - \bar{\hat{\theta}})}{S_{\hat{\theta}}},$$

where $\bar{\hat{\theta}}$ is the mean of $\hat{\theta}_j$ and $S_{\hat{\theta}}$ is the standard deviation of $\hat{\theta}_j$. Item parameter estimates in terms of $\hat{a}_i$ and $\hat{b}_i$ are also adjusted as

$$\hat{b}_i^* = \frac{(\hat{b}_i - \bar{\hat{\theta}})}{S_{\hat{\theta}}} \quad \text{and} \quad \hat{a}_i^* = S_{\hat{\theta}} \hat{a}_i .$$

Rescale those new θ s (estimated in the step 2) in Step 3, i.e., finishing one cycle of the paradigm so that they correspond to the particular metric. After this transformation, go back and repeat the Steps 1-3 until the convergence criterion is met.

This procedure continues until the overall log-likelihood does not change much from stage to stage.

Some limitations of the JML approach:

- Ability parameter estimates for perfect (all items answered correctly) and zero (all items answered incorrectly) examinee response patterns do not exist.
- Item-parameter estimates for perfect (all examinees answer correctly) and zero (all examinees answer incorrectly) item response patterns do not exist.
- In the 2PL and 3PL model estimation may fail. For example, items that have very poor discrimination sometimes result in difficulty estimates that move to infinity. To handle this problem, bounds are sometimes set on parameter values, or the offending items (or examinees) are removed from the analysis.
- Heywood cases – Discrimination parameters for some items may move to infinity. This problem is the same problem that occurs in factor analysis when estimates of the uniqueness are 0 or negative.

Definitions of *consistent, efficient,* and *sufficient* estimators

- An estimator $\hat{\mu}$ is a consistent estimator of μ provided that as the sample size increases, the estimate converges to the true parameter value with probability 1, or $P(\hat{\mu} \to \mu) \to 1$ as $N \to \infty$.
- An estimator $\hat{\mu}$ is an efficient estimator of μ if it is unbiased and produces the minimum variance possible for an unbiased estimator of μ.
- An estimator $\hat{\mu}$ is a sufficient estimator for μ provided once $\hat{\mu}$ is known, there is no additional information in the data concerning μ, or $P(U \mid \hat{\mu}, \mu) = P(U \mid \hat{\mu})$.

Definitions of *incidental* and *structural* parameters

Suppose $X_1, X_2, \ldots, X_j$ are independently distributed random variables whose distribution is dependent on the parameters τ and θ_j. Because the τ stays constant across j, they are regarded as *structural* parameters, while the θ_j, which vary over j are *incidental* parameters. In IRT, item parameters are structural parameters while ability parameters are incidental parameters.

- In JML, item parameter estimates can not be theoretically proven to be consistent (for the 2PL and 3PL models). This is because as sample size increases, so do the number of unknown ability parameters. (However, Swaminathan & Gifford (1983) have shown empirically that consistent estimates result when both the number of items and the number of examinees become large). The answer to this problem is to try to find a way to remove the incidental parameters from the analysis. Conditional maximum likelihood (CML) procedure does this by using a sufficient statistic for the θ s. Marginal maximum likelihood (MML) procedure does this by integrating the θ s out of the log-likelihood function.

Quality of Item Parameter Estimates obtained using JML

- Bias of the estimates: Using the 3PL, Lord (1983) found that easy and medium difficulty items have negative bias in the difficulty parameter, while difficult items had a positive bias. The bias in item discriminations was always positive. Bias in guessing parameter was always negative.

- Standard errors: Thissen & Wainer (1982) derives asymptotic standard errors of 1PL, 2PL, and 3PL assuming abilities of examinees are known. For the 3PL, the standard errors are only small in the middle ranges of α and β and when c is small. According to Thissen & Wainer (1982), the large standard errors for the 3PL make it inferior to the 1PL or 2PL.

$$
\begin{bmatrix} \hat{\zeta}_1 \\ \hat{\lambda}_1 \\ \hat{\zeta}_2 \\ \hat{\lambda}_2 \\ \vdots \\ \hat{\zeta}_n \\ \hat{\lambda}_n \\ \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_N \end{bmatrix}_{t+1} = \begin{bmatrix} \hat{\zeta}_1 \\ \hat{\lambda}_1 \\ \hat{\zeta}_2 \\ \hat{\lambda}_2 \\ \vdots \\ \hat{\zeta}_n \\ \hat{\lambda}_n \\ \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_N \end{bmatrix}_t - \begin{bmatrix} \frac{\partial^2 L}{\partial \zeta_1^2} & \frac{\partial^2 L}{\partial \zeta_1 \partial \lambda_1} & & & & & & & & & \\ \frac{\partial^2 L}{\partial \lambda_1 \partial \zeta_1} & \frac{\partial^2 L}{\partial \lambda_1^2} & & & & & & & & & \\ & & \frac{\partial^2 L}{\partial \zeta_2^2} & \frac{\partial^2 L}{\partial \zeta_2 \partial \lambda_2} & & & & & & & \\ & & \frac{\partial^2 L}{\partial \lambda_2 \partial \zeta_2} & \frac{\partial^2 L}{\partial \lambda_2^2} & & & & & & & \\ & & & & \ddots & & & & & & \\ & & & & & \frac{\partial^2 L}{\partial \zeta_n^2} & \frac{\partial^2 L}{\partial \zeta_n \partial \lambda_n} & & & & \\ & & & & & \frac{\partial^2 L}{\partial \lambda_n \partial \zeta_n} & \frac{\partial^2 L}{\partial \lambda_n^2} & & & & \\ & & & & & & & \frac{\partial^2 L}{\partial \theta_1^2} & & & \\ & & & & & & & & \frac{\partial^2 L}{\partial \theta_2^2} & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & \frac{\partial^2 L}{\partial \theta_N^2} \end{bmatrix}_t^{-1} \times \begin{bmatrix} \frac{\partial L}{\partial \zeta_1} \\ \frac{\partial L}{\partial \lambda_1} \\ \frac{\partial L}{\partial \zeta_2} \\ \frac{\partial L}{\partial \lambda_2} \\ \vdots \\ \frac{\partial L}{\partial \zeta_n} \\ \frac{\partial L}{\partial \lambda_n} \\ \frac{\partial L}{\partial \theta_1} \\ \frac{\partial L}{\partial \theta_2} \\ \vdots \\ \frac{\partial L}{\partial \theta_N} \end{bmatrix}_t
$$