IRT Parameter Estimation

Marginal Maximum Likelihood in IRT

Steven W. Nydick

University of Minnesota

May 23, 2012

Outline

Notation

$j = 1, 2, \ldots, J$	Item Enumeration
$i = 1, 2, \ldots, n$	Examinee Enumeration
$k = 1, 2, \ldots, q$	Group Enumeration
$y_{ij} = \{1, 0\}$	Response of examinee i to item j
$\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})$	Response vector of examinee i
a_j, b_j, c_j	Parameters of item j
$\boldsymbol{\phi}_j, \boldsymbol{\Phi}$	Vector/Matrix of "true" item parameters
$\boldsymbol{\Gamma}$	Parameters of population ability distribution
θ_i	True ability of examinee i
X_k	True ability of group k
n_{kj}	Number of attempts by group k to item j
r_{kj}	Number of correct by group k to item j

$$P_j(\theta_i) = \Pr(y_{ij} = 1 | \theta_i, \boldsymbol{\phi}_j) \qquad Q_j(\theta_i) = 1 - P_j(\theta_i) = \Pr(y_{ij} = 0 | \theta_i, \boldsymbol{\phi}_j)$$

A Brief History

Maximum likelihood estimation was introduced by R.A. Fisher in 1912.

Reasons we use it in statistics include the following.

1. We want to estimate distribution parameters.
2. MLE is an Intuitive method of estimation.
3. MLE typically results in consistent estimate of parameters.

MLE is the most widely used estimation method in statistics.

Likelihood: An Outline

First, define a likelihood function.

1. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a response vector from distribution $f(x|\mathbf{\Gamma})$.
   - $\mathbf{\Gamma}$ are the parameters of the distribution.
2. Then a "probability density function" $f_n(\mathbf{x}|\mathbf{\Gamma})$ provides the "likelihood" of observing $\mathbf{\Gamma}$.

If we *know* $\mathbf{\Gamma}$ but *do not know* $\mathbf{x}$, then

$$f[(x_1, x_2)|\mathbf{\Gamma}] = f(x_1|\mathbf{\Gamma}) \cdot f(x_2|\mathbf{\Gamma})$$

assuming independence.

Probability of a Vector

And assuming independence of all responses, we have

$$f(\mathbf{x} = (x_1, x_2, \ldots, x_n)|\mathbf{\Gamma}) = \prod_{i=1}^{n} f(x_i|\mathbf{\Gamma}) = f_n(\mathbf{x}|\mathbf{\Gamma})$$

This is the "probability" of observing *a vector*, so order matters.

The trick to likelihood inference is turning the pdf on its head:

Assume that we have a density distribution where we know $\boldsymbol{x}$ *and we are looking at the likelihood of observing the parameters.*

Probability: events that havent happened yet
Likelihood: events that have already happened

The Maximum Likelihood

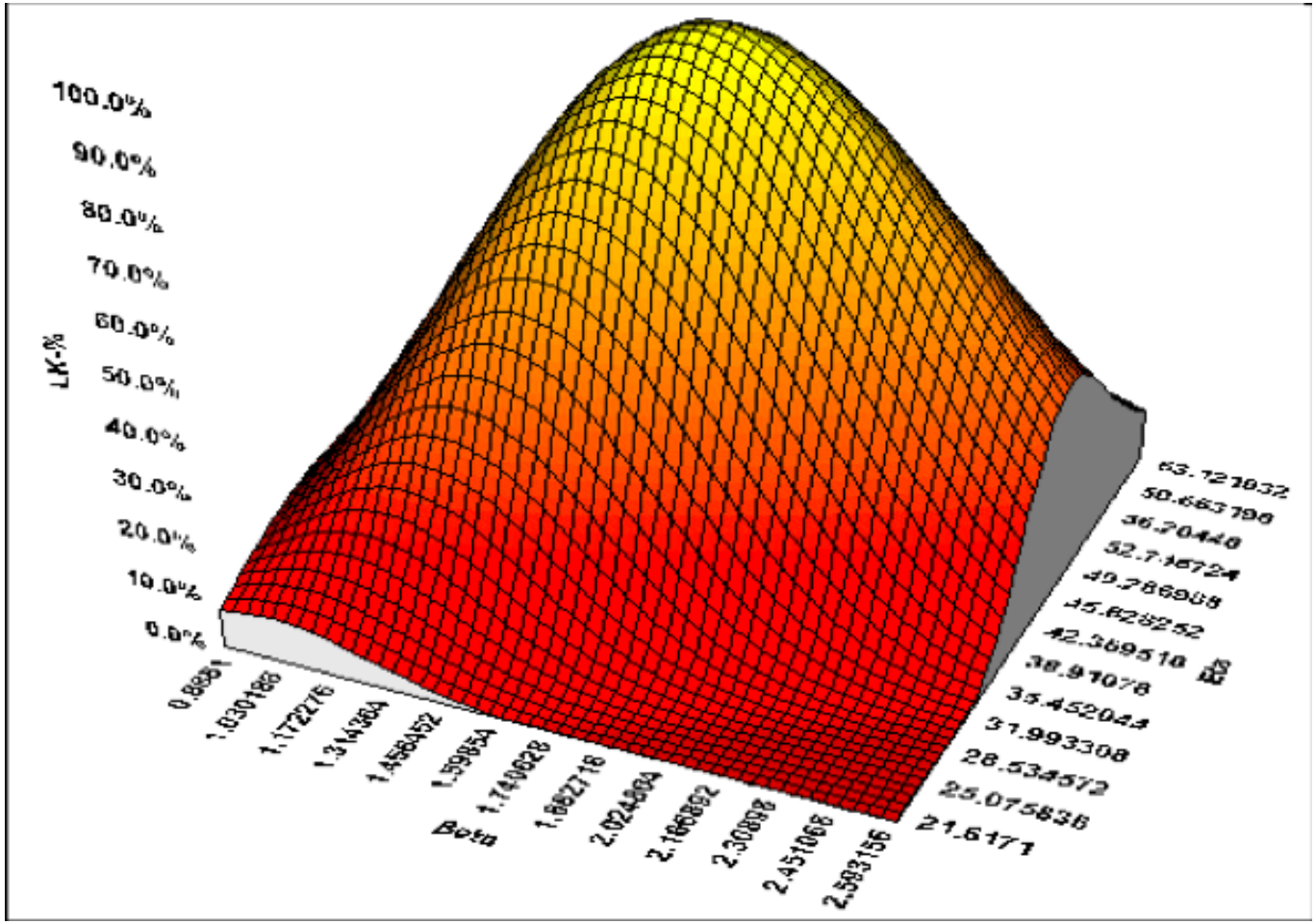
How can we estimate the parameters after knowing $\mathbf{x}$?

The *maximum* of the likelihood might be a good idea. Why?

1. The simple answer: Most of the stuff is in the small space surrounding the maximum of the likelihood.
   - Stuff indicates "likelihood of parameters."
2. The complex answer: The MLE is generally consistent, efficient, and asymptotically normal.

The Maximum Likelihood: A Visual

The following is an example of a likelihood function with two variables.



The Maximum Likelihood: A Method

So, solve the following equation.

$$\frac{d[f_n(\mathbf{x}|\mathbf{\Gamma})]}{d\mathbf{\Gamma}} = \frac{d[\prod_{i=1}^{n} f(x_i|\mathbf{\Gamma})]}{d\mathbf{\Gamma}} = \mathbf{0}$$

... which will be a gradient if there is more than one parameter.

But a bunch of product terms is a pain.

However, $\ln(x) = \log(x)$ is a really nice function:

1. $\ln(x)$ is monotonically increasing, so the maximum doesn't change.
2. $\ln(x_1 \cdot x_2) = \ln(x_1) + \ln(x_2)$

The Maximum Likelihood: A Method

So, solve the following equation.

$$\frac{d[L(\mathbf{x}|\mathbf{\Gamma})]}{d\mathbf{\Gamma}} = \frac{\sum_{i=1}^{n} d[\log(f(x_i|\mathbf{\Gamma}))]}{d\mathbf{\Gamma}} = \mathbf{0}$$

... which is a nicer gradient, both for you and your friend the computer.

Finding our Function

Now we want to apply maximum likelihood to IRT.

First, define the probability mass function (PMF)

$$f_3(y_{ij}) = \begin{cases} P_j(\theta_i) & \text{if } y_{ij} = 1 \\ Q_j(\theta_i) & \text{if } y_{ij} = 0 \end{cases}$$

... which is a simple Bernoulli r.v. for a given θ and a given item.

Often, the probability of response is assumed to hide a logit link.

The Logistic Functions

The Bernoulli r.v. maps latent abilities to observed responses.

For the 3PL model, the probability of a correct response is

$$\begin{aligned} P_j(\theta_i) &= c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} \\ &= c_j + (1 - c_j)\frac{1}{1 + \exp[-a_j(\theta_i - b_j)]} \end{aligned}$$

And the probability of an incorrect response is

$$\begin{aligned} Q_j(\theta_i) = 1 - P_j(\theta_i) &= 1 - \left[c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}\right] \\ &= (1 - c_j) - (1 - c_j)\left[\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}\right] \\ &= (1 - c_j)\left[1 - \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}\right] \end{aligned}$$

The Logistic Functions: 2

To make our lives easier, define a 2PL model.

$$f_2(y_{ij}) = \begin{cases} P_j^\star(\theta_i) & \text{if } y_{ij} = 1 \\ Q_j^\star(\theta_i) & \text{if } y_{ij} = 0 \end{cases}$$

The 2PL model implies

$$P_j^\star(\theta_i) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}$$

and

$$\begin{aligned} Q_j^\star(\theta_j) = 1 - \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} &= \frac{1}{1 + \exp[a_j(\theta_i - b_j)]} \\ &= \frac{\exp[-a_j(\theta_i - b_j)]}{1 + \exp[-a_j(\theta_i - b_j)]} \end{aligned}$$

The Logistic Functions 2.2 ... 3.2 ... ?

The 3PL probabilities relate to the 2PL probabilities.

$$\begin{aligned} Q_j(\theta_i) &= (1 - c_j)\left[1 - \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}\right] \\ &= (1 - c_j)[1 - P_j^\star(\theta_i)] \\ &= (1 - c_j)[Q_j^\star(\theta_i)] \end{aligned}$$

- It is silly to define both functions as PMFs.
- We are implying that both the 3PL model and 2PL model hold at the same time.
- Our 2PL model definition is just for computational simplicity.

A Few More Things

Next, we want to note that

$$f_3(y_{ij}) = \begin{cases} P_j(\theta_i) & \text{if } y_{ij} = 1 \\ Q_j(\theta_i) & \text{if } y_{ij} = 0 \end{cases}$$

can be written more compactly as

$$f_3[y_{ij}|\theta_i, \boldsymbol{\phi}_j = (a_j, b_j, c_j)] = P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{(1-y_{ij})}$$

Remember that y_{ij} is a Bernoulli random variable, so

$$y_{ij} \in \{0, 1\}$$

A Few More Things

A major assumption in IRT:

Conditional on the trait (θ_i), the observations are independent.

Note that responses are not i.i.d. unless the same item is given to the same person (by using memory erasing capabilities).

Each item/person combination is a **different** Bernoulli random variable, and assuming independence of responses and ordered items for a given person, we have

$$f_3(\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})|\theta_i, \mathbf{\Phi}) = \prod_{j=1}^{J} P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{(1-y_{ij})}$$

A Few More Things

Another major assumption in IRT:

> *A given person's response vector $\boldsymbol{y}_i$ is independent from all other response vectors after taking the trait into consideration.*

And assuming ordered items and ordered persons, we have

$$f_3\left(\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n\end{bmatrix} | \boldsymbol{\theta}, \boldsymbol{\Phi}\right) = \prod_{i=1}^{n}\prod_{j=1}^{J} P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{(1-y_{ij})}$$

MLE in IRT

For now we assume that

1. we have the response matrix,
2. we know the vector of $\boldsymbol{\theta}$, and
3. we have the likelihood of observing $\mathbf{Y}$ given a variety of $\boldsymbol{\Phi}$.

And just as in the earlier picture

> *We want to find our best guess of the location of the parameters. And using Fisherian theory, our best guess, and the one that has the best properties, is as the maximum of the likelihood.*

- (Think of weird likelihood picture in a lot of dimensions.)
- (Or don't ...)

MLE in IRT

Therefore, maximize

$$f_3\left(\mathbf{Y}|\boldsymbol{\theta},\,\boldsymbol{\Phi}\right) = \prod_{i=1}^{n}\prod_{j=1}^{J} P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{(1-y_{ij})}$$

which is equivalent to maximizing

$$\begin{aligned} L(\mathbf{Y}|\boldsymbol{\theta},\,\boldsymbol{\Phi}) &= \log\left(\prod_{i=1}^{n}\prod_{j=1}^{J} P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{(1-y_{ij})}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\Big(y_{ij}\log P_j(\theta_i) + (1-y_{ij})\log Q_j(\theta_i)\Big) \end{aligned}$$

Results of the Derivation

See Harwell, Baker, and Zwarts (1988) for the details of the derivation.

Then letting

$$\omega_{ij} = \frac{P_j^\star(\theta_i) Q_j^\star(\theta_i)}{P_j(\theta_i) Q_j(\theta_i)}$$

the derivative of the log-likelihood for a given item j is

$$\frac{\partial L}{\partial a_j} = (1 - c_j) \sum_{i=1}^{n} \Big([y_{ij} - P_j(\theta_i)] \cdot \omega_{ij} (\theta_i - b_j) \Big)$$

$$\frac{\partial L}{\partial b_j} = (1 - c_j)(-a_j) \sum_{i=1}^{n} \Big([y_{ij} - P_j(\theta_i)] \cdot \omega_{ij} \Big)$$

$$\frac{\partial L}{\partial c_j} = (1 - c_j)^{-1} \sum_{i=1}^{n} \left[\frac{y_{ij} - P_j(\theta_i)}{P_j(\theta_i)} \right]$$

Results of the Derivation

And (for each item) we have three equations and three unknowns.

Finally, set

$$\begin{pmatrix} \frac{\partial L}{\partial a_j} \\ \frac{\partial L}{\partial b_j} \\ \frac{\partial L}{\partial c_j} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

to search for a maximum, iterate, and solve.

Another Way

We can also group examinees at a finite number of ability levels.

Let

- n_{kj} be the number of examinees at level k who response to item j,
- r_{kj} be the number of examinees who *correctly* respond to item j at ability level k, and
- $\hat{p}_{kj} = \frac{r_{kj}}{n_{kj}}$ be the estimated probability of response to item j at ability level k.

And instead of a single Bernoulli response for each examinee/item, we have a sum of Bernoulli responses for each group/item combination.

Another Distribution

Assume that we *know* p_j for each group on item j.

$$f_4(r_{kj}) = \begin{cases} p_{1kj} & \text{if } r_{kj} = 1 \\ p_{2kj} & \text{if } r_{kj} = 2 \\ \vdots & \vdots \\ p_{nkj} & \text{if } r_{kj} = n_{kj} \end{cases}$$

How do we write f_4 compactly?

$$f_4[r_{kj}|X_k, n_{kj}, \boldsymbol{\phi}_j] = \binom{n_{kj}}{r_{kj}} P_j(X_k)^{r_{kj}} Q_j(X_k)^{(n_{kj}-r_{kj})}$$

Another Likelihood

Assuming X_k is known for each k, r_{kj} and n_{kj} are observed at each ability level, then the likelihood is as follows.

$$f_4[\mathbf{r}|\mathbf{X},\, \mathbf{n},\, \mathbf{\Phi}] = \prod_{k=1}^{q}\prod_{j=1}^{J}\binom{n_{kj}}{r_{kj}} P_j(X_k)^{r_{kj}} Q_j(X_k)^{(n_{kj}-r_{kj})}$$

And the loglikelihood is as follows.

$$L_4[\mathbf{r}|\mathbf{X},\, \mathbf{n},\, \mathbf{\Phi}] = m + \sum_{k=1}^{q}\sum_{j=1}^{J}\Big[r_{kj}\log P_j(X_k) + (n_{kj}-r_{kj})\log Q_j(X_k)\Big]$$

Another Openin' Another Show

And, by analogy of the previous "derivation," for a given item j,

$$\frac{\partial L}{\partial a_j} = (1 - c_j) \sum_{k=1}^{q} \Big([r_{kj} - n_{kj} \cdot P_j(X_k)] \cdot \omega_{kj} (X_k - b_j) \Big)$$

$$\frac{\partial L}{\partial b_j} = (1 - c_j)(-a_j) \sum_{k=1}^{q} \Big([r_{kj} - n_{kj} \cdot P_j(X_k)] \cdot \omega_{kj} \Big)$$

$$\frac{\partial L}{\partial c_j} = (1 - c_j)^{-1} \sum_{k=1}^{q} \left[\frac{r_{kj} - n_{kj} \cdot P_j(X_k)}{P_j(X_k)} \right]$$

Again -- three equations and three unknowns.

Set equal to 0, iterate, and solve.

Simplified JMLE

We generally do not know $\boldsymbol{\theta}$ or $\mathbf{X}$. What do we do?

1. Use standardized raw test scores as initial "known ability values", or
2. assume fixed ability points using raw-test scores to "group" examinees.
3. Solve for item parameters individually (item-by-item).
4. Re-estimate $\boldsymbol{\theta}/\mathbf{X}$ for each person or fixed ability point to group examinees.
5. Anchor $\boldsymbol{\theta}/\mathbf{X}$ by standardizing (i.e. converting to z-scores).
6. Re-estimate item parameters individually (item-by-item).
7. "Ping-Pong" until some convergence criterion is met.

The above procedure is called "JMLE" and implemented in LOGIST.

A Metaphor

Calibration in Pictures

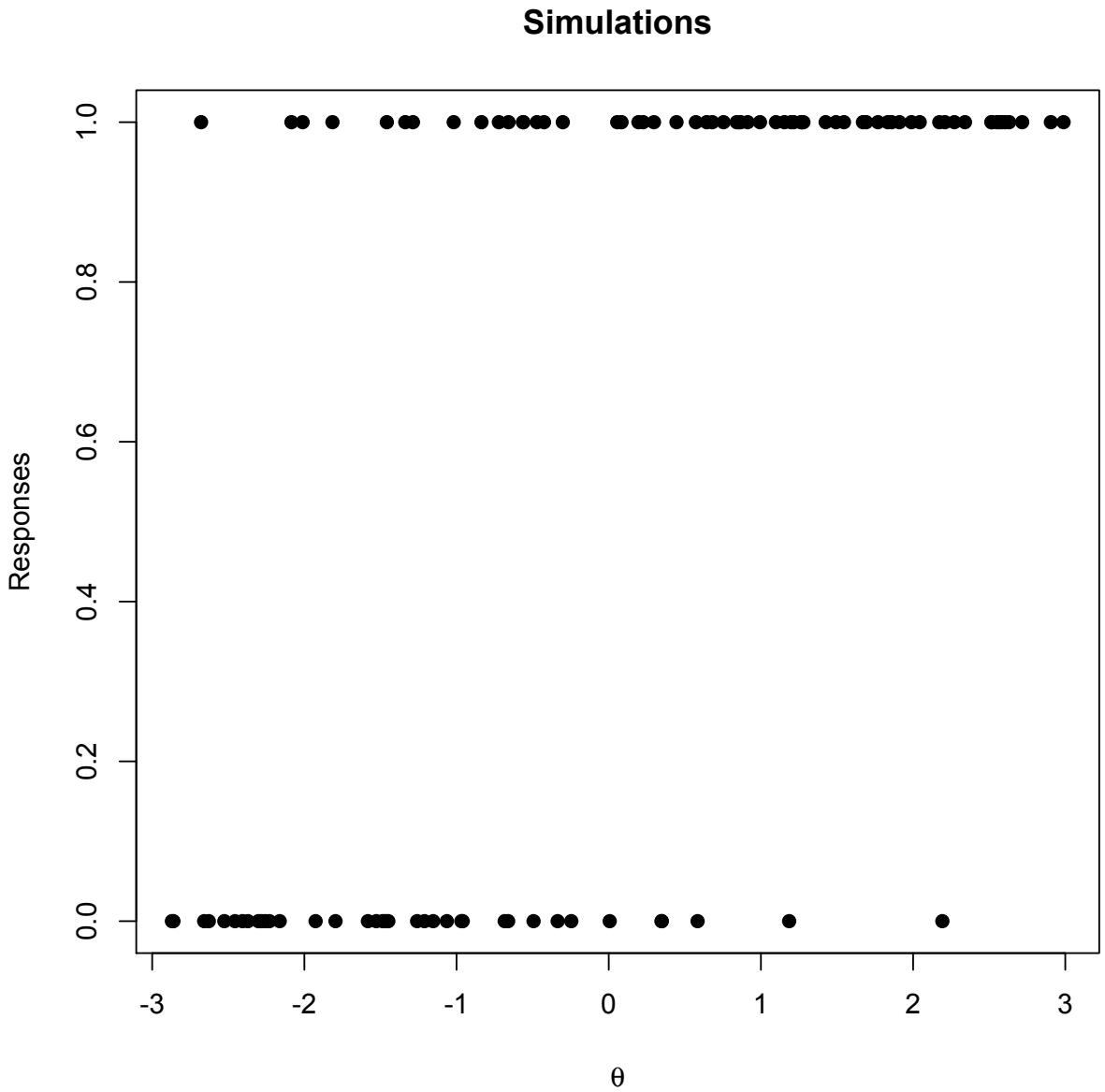
Using the first set of equations

> *... there will be a bunch of 0s and 1s above and below the theoretical ogive.*

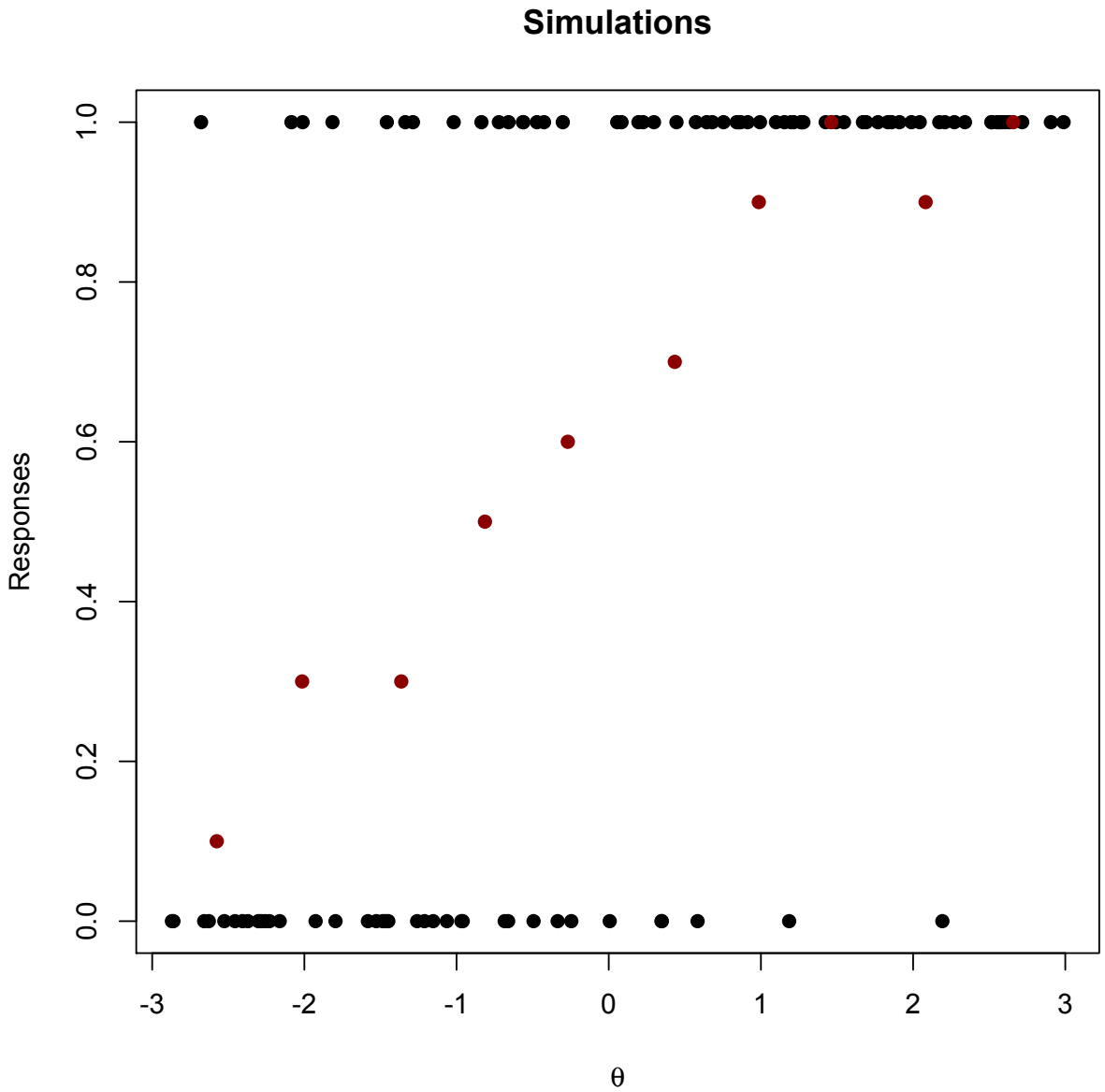
Using the second set of equations

> *... there will be an estimate probability at each of the k levels.*

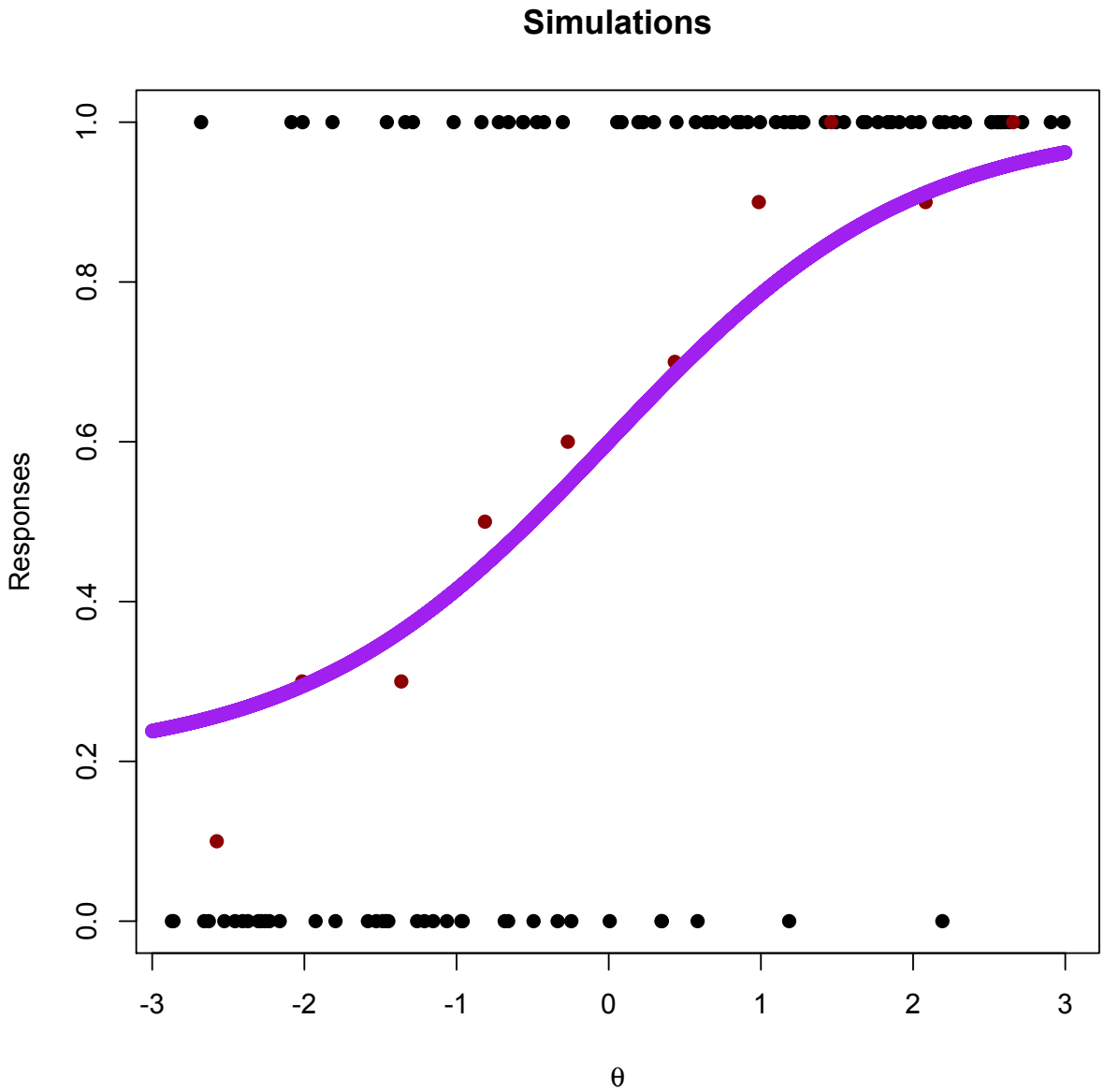
Calibration in Pictures



Calibration in Pictures



Calibration in Pictures



The End of JMLE

There is a major problem with JMLE.

> *We must assume that ability is known/fixed to estimate items, and we must assume that item parmaeters are known to estimate ability.*

The first set of JMLE equations are not even consistent!

Why cannot we just ignore $\boldsymbol{\theta}$ in the estimation?

JMLE and MMLE

A simple comparison of JMLE to MMLE:

1. JMLE is a fully-fixed effects model, in that both item parameters and ability are fixed parameters to be estimated.
   - Assuming fixed-effects force us to estimate more parameters than desired and eliminates the benefit of consistency.
2. MMLE is a mixed-effects model, with item parameters as fixed effects and ability parameters as random effects.
   - Assuming random effects allows us to posit a distribution and remove them from the estimating equations.

The Beginning: Bayes

Using a distribution of θ requires Bayes theorem.

$$f_5(\theta|\mathbf{y}_i, \mathbf{\Phi}, \mathbf{\Gamma}) = \frac{f_3(\mathbf{y}_i|\theta, \mathbf{\Phi})g(\theta|\mathbf{\Gamma})}{\int f_3(\mathbf{y}_i|\theta, \mathbf{\Phi})g(\theta|\mathbf{\Gamma})d\theta}$$
$$= \frac{f_3(\mathbf{y}_i|\theta, \mathbf{\Phi})g(\theta|\mathbf{\Gamma})}{f_6(\mathbf{y}_i|\mathbf{\Phi})}$$

We want to find the distribution of

$$f_6(\mathbf{y}_i|\mathbf{\Phi}) = \int f_3(\mathbf{y}_i|\theta, \mathbf{\Phi})g(\theta|\mathbf{\Gamma})d\theta$$

... by integrating out θ to obtain a marginal distribution of $\mathbf{\Gamma}$.

The MMLE Likelihood: B & L

How do we obtain this "prior" distribution of θ?

If we *have* a prior distribution, then maximize

$$f_6(\mathbf{Y}|\mathbf{\Phi}) = \prod_{i=1}^{n} f_6(\mathbf{y}_i|\mathbf{\Phi})$$

which is equivalent to maximizing

$$L(\mathbf{Y}|\mathbf{\Phi}) = \sum_{i=1}^{n} \log f_6(\mathbf{y}_i|\mathbf{\Phi})$$

The MMLE Likelihood: B & L

Why do we want to maximize the previous equations?

Given: The response vectors for each person.

Unknown: The ability associated with those response vectors.

By eliminating the irritating thing of having to know each persons ability, then we can estimate the item parameters directly.

Results of the Derivation: B & L

See Harwell et al. (1998) for the details of the derivation.

For a given item j,

$$\frac{\partial L}{\partial a_j} = (1 - c_j) \sum_{i=1}^{n} \int \Big([y_{ij} - P_j(\theta)] \cdot \omega_{ij} (\theta - b_j) \Big) f_5(\theta | \mathbf{y}_i, \mathbf{\Phi}, \mathbf{\Gamma}) d\theta$$

$$\frac{\partial L}{\partial b_j} = (1 - c_j)(-a_j) \sum_{i=1}^{n} \int \Big([y_{ij} - P_j(\theta)] \cdot \omega_{ij} \Big) f_5(\theta | \mathbf{y}_i, \mathbf{\Phi}, \mathbf{\Gamma}) d\theta$$

$$\frac{\partial L}{\partial c_j} = (1 - c_j)^{-1} \sum_{i=1}^{n} \int \left[\frac{y_{ij} - P_j(\theta)}{P_j(\theta)} \right] f_5(\theta | \mathbf{y}_i, \mathbf{\Phi}, \mathbf{\Gamma}) d\theta$$

Results of the Derivation: B & L

The MML derivation

1. looks similar to the MLE/JMLE derivation, but
2. the distribution of θ is integrated out of the equation.

Of course the only additional part to MMLE is the distribution of θ.

Quadrature

How do we find the perform the integral inside the gradient?

If a distribution is continuous and has *finite* moments, it can be approximated to any desired degree of accuracy with a histogram.

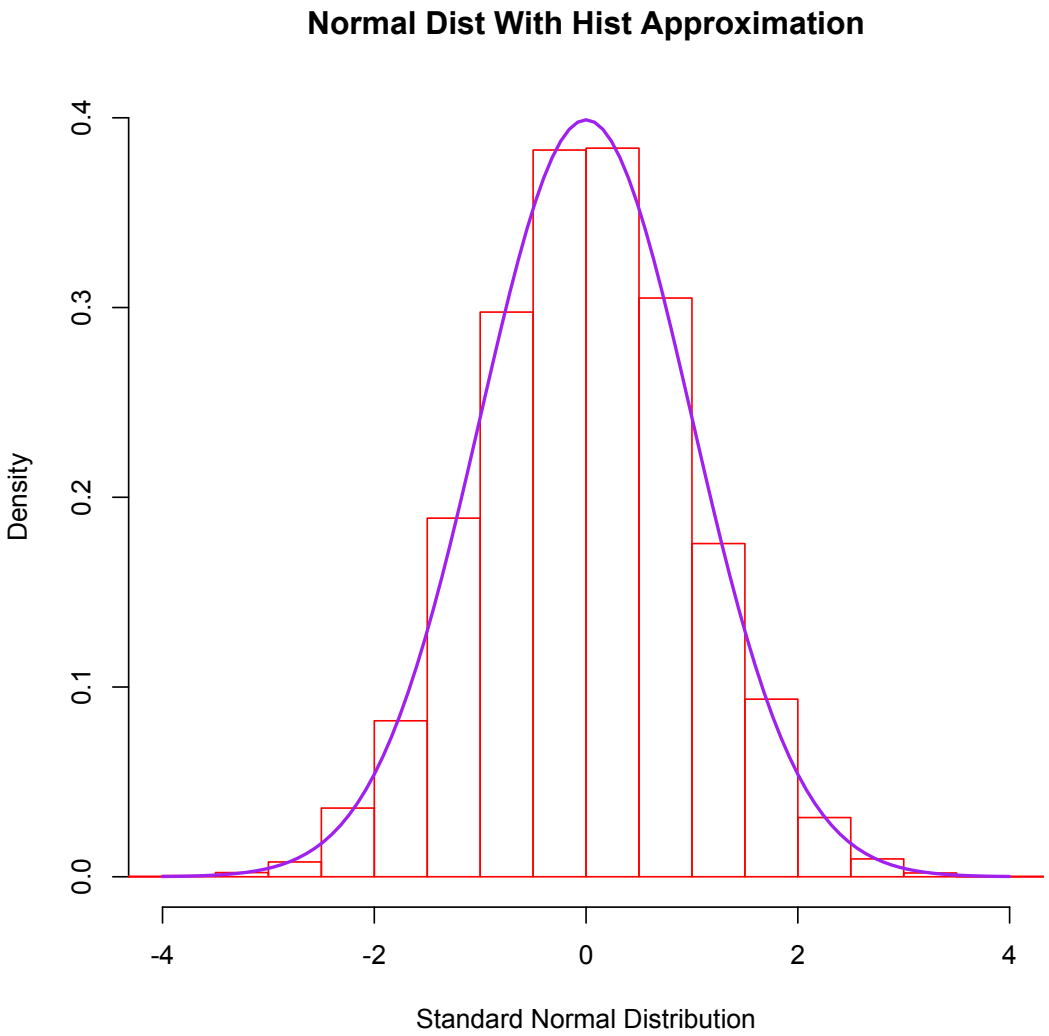
Therefore, we must assume that

$$g(\theta|\mathbf{\Gamma})$$

is continuous.

A usual assumption is that $g(\theta|\mathbf{\Gamma})$ is normally distributed.

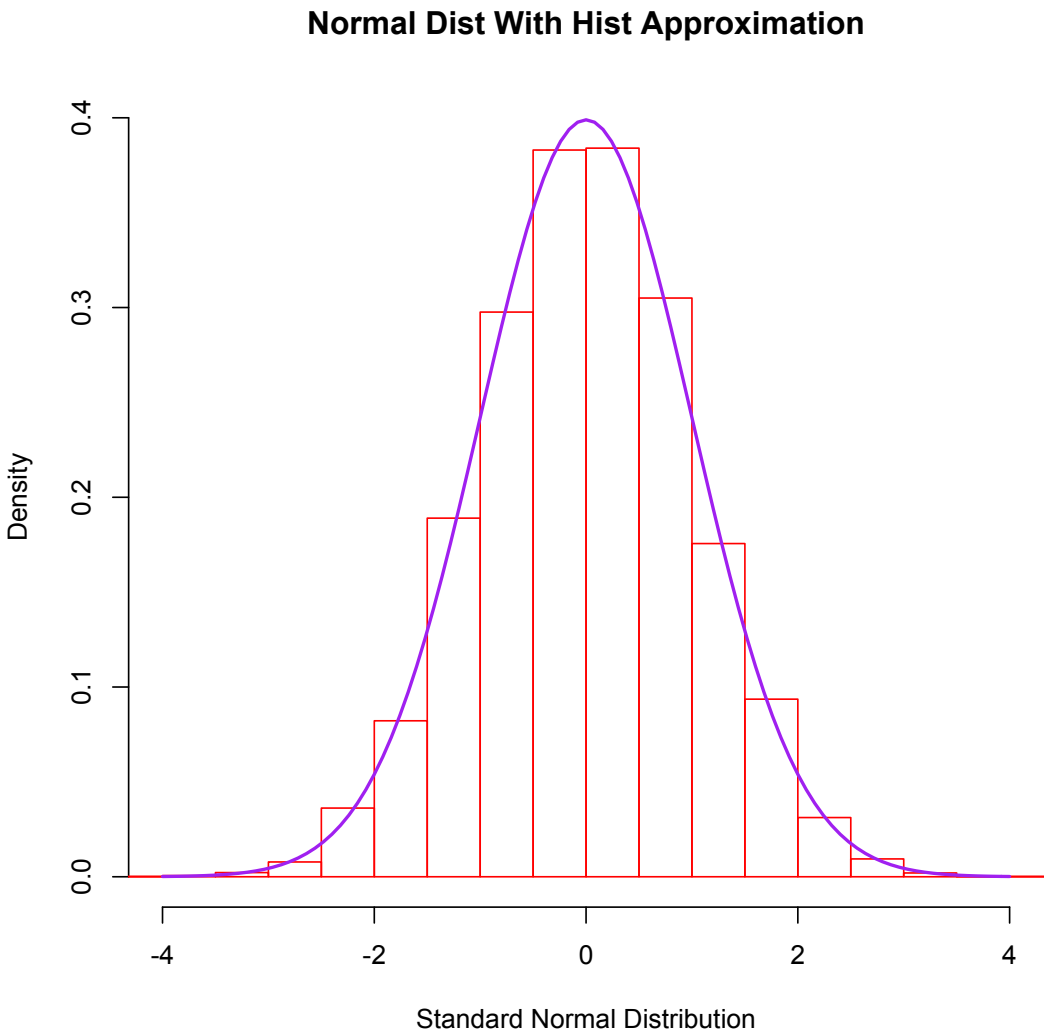
Quadrature



Define

1. X_k as the rectangle midpoint ("node").
2. $A(X_k)$ as the weight given the node.

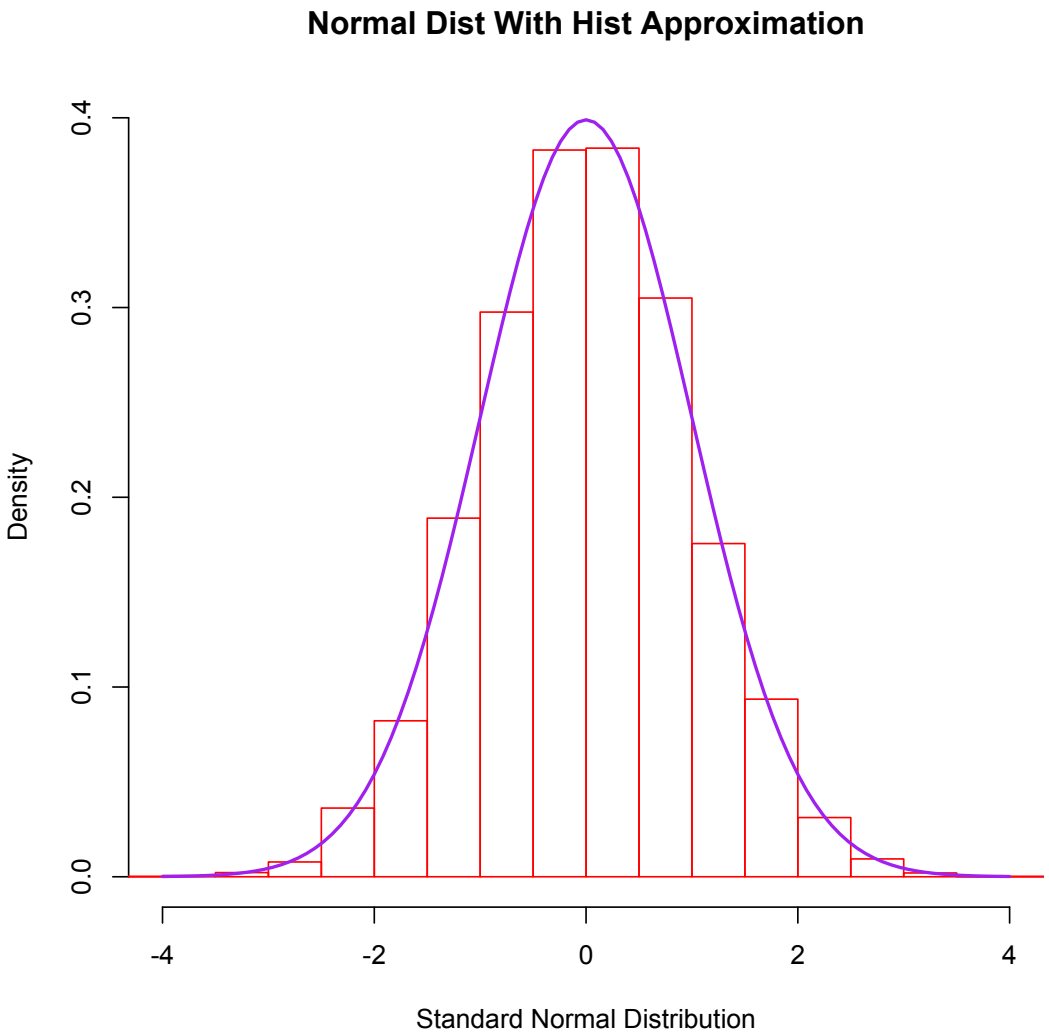
Quadrature



Quadrature is Riemann calculus ... backwards.

- The number of nodes is finite ($k = 1, \ldots, q$).
- Because the number of nodes is finite, the distribution is usually bounded.

Quadrature



Quadrature is Riemann calculus ... backwards.

1. Equally spaced intervals $\rightarrow$ Easier computation
2. Non-equally space intervals $\rightarrow$ Improved loss function for equivalent number of bins

Implementing Quadrature

$g(\theta|\mathbf{\Gamma})$ can be determined by empirical methods.

Now rewrite our equation taking into consideration quadrature.

We are approximating θ with discrete values, so if θ is in a specific "bin", k, we then approximate θ with X_k.

A Finite Bayes

Therefore,

$$f_5(\theta|\mathbf{y}_i,\, \mathbf{\Phi},\, \mathbf{\Gamma}) = \frac{f_3(\mathbf{y}_i|\theta,\, \mathbf{\Phi})g(\theta|\mathbf{\Gamma})}{\int f_3(\mathbf{y}_i|\theta,\, \mathbf{\Phi})g(\theta|\mathbf{\Gamma})d\theta}$$

becomes

$$f_5(X_k|\mathbf{y}_i,\, \mathbf{\Phi},\, \mathbf{\Gamma}) = \frac{\prod_{j=1}^{J} P_j(X_k)^{y_{ij}} Q_j(X_k)^{(1-y_{ij})} A(X_k)}{\sum_{k=1}^{q} \prod_{j=1}^{J} P_j(X_k)^{y_{ij}} Q_j(X_k)^{(1-y_{ij})} A(X_k)}$$

where

$$g(\theta|\mathbf{\Gamma}) \approx A(X_k) \qquad f_3(\mathbf{y}_i|\theta,\, \mathbf{\Phi}) \approx P_j(X_k)^{y_{ij}} Q_j(X_k)^{(1-y_{ij})}$$

Bock and Lieberman MMLE Solution

Now X_k takes on finite values, and $A(X_k)$ are the corresponding weights.

Adjusting the solutions for each of the parameters results in

$$\frac{\partial L}{\partial a_j} = (1 - c_j) \sum_{i=1}^{n} \sum_{k=1}^{q} \Big([y_{ij} - P_j(X_k)] \cdot \omega_{kj} (X_k - b_j) \Big) f_5(X_k | \mathbf{y}_i, \, \mathbf{\Phi}, \, \mathbf{\Gamma})$$

$$\frac{\partial L}{\partial b_j} = (1 - c_j)(-a_j) \sum_{i=1}^{n} \sum_{k=1}^{q} \Big([y_{ij} - P_j(X_k)] \cdot \omega_{kj} \Big) f_5(X_k | \mathbf{y}_i, \, \mathbf{\Phi}, \, \mathbf{\Gamma})$$

$$\frac{\partial L}{\partial c_j} = (1 - c_j)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{q} \left[\frac{y_{ij} - P_j(X_k)}{P_j(X_k)} \right] f_5(X_k | \mathbf{y}_i, \, \mathbf{\Phi}, \, \mathbf{\Gamma})$$

Bock and Aitken Transition

Technically:

- You can set the gradient equal to 0 and solve through approximation or Newton-Raphson.

However:

- The quadrature function is conditional on *all* of the parameters.

Bock and Aitken proposed an alternative solution...

Bock and Aitken Notation

Step One:

$$E[n_k] : \bar{n}_k = \sum_{i=1}^{n} [f_5(X_k|\mathbf{y}_i, \mathbf{\Phi}, \mathbf{\Gamma})]$$
$$= \sum_{i=1}^{n} \left[\frac{\prod_{j=1}^{J} P_j(X_k)^{y_{ij}} Q_j(X_k)^{(1-y_{ij})} A(X_k)}{\sum_{k=1}^{q} \prod_{j=1}^{J} P_j(X_k)^{y_{ij}} Q_j(X_k)^{(1-y_{ij})} A(X_k)} \right]$$

The equation is simplier than it appears:

- $f_5(X_k|\mathbf{y}_i, \mathbf{\Phi}, \mathbf{\Gamma})$ is essentially the "probability of being in any group" given a person's response vector.
- Prob of being in a group summed over all response vectors is the "expected number" in that group.

Bock and Aitken Notation

Step Two:

$$E[r_{kj}] : \bar{r}_{kj} = \sum_{i=1}^{n} y_{ij}[f_5(X_k|\mathbf{y}_i, \mathbf{\Phi}, \mathbf{\Gamma})]$$

$$= \sum_{i=1}^{n} y_{ij} \left[\frac{\prod_{j=1}^{J} P_j(X_k)^{y_{ij}} Q_j(X_k)^{(1-y_{ij})} A(X_k)}{\sum_{k=1}^{q} \prod_{j=1}^{J} P_j(X_k)^{y_{ij}} Q_j(X_k)^{(1-y_{ij})} A(X_k)} \right]$$

The equation is simplier than it appears:

- $f_5(X_k|\mathbf{y}_i, \mathbf{\Phi}, \mathbf{\Gamma})$ is a probability of being in a particular category for reponse vector i.
- y_{ij} is the actual response of person i
- Prob of being in a category multiplied by the response summed over all people is the expected number of correct responses at a given X_k.

Bock and Aitken MMLE Solution

Adjusting the Bock and Lieberman solution:

$$\frac{\partial L}{\partial a_j} = (1 - c_j) \sum_{k=1}^{q} \left([\bar{r}_{kj} - \bar{n}_k P_j(X_k)] \cdot \omega_{kj} (X_k - b_j) \right)$$

$$\frac{\partial L}{\partial b_j} = (1 - c_j)(-a_j) \sum_{k=1}^{q} \left([\bar{r}_{kj} - \bar{n}_k P_j(X_k)] \cdot \omega_{kj} \right)$$

$$\frac{\partial L}{\partial c_j} = (1 - c_j)^{-1} \sum_{k=1}^{q} \left[\frac{\bar{r}_{kj} - \bar{n}_k P_j(X_k)}{P_j(X_k)} \right]$$

What do the above equations mean? We have "expected number of correct responses" and "probability of a correct response times the expected number of people who have that probability."

Solving our Equations

To solve the equations we

1. need provisional item parameter estimates,
2. use weights and quadrature points to compute posterior probability ($f_5(X_k)$) for each examinee,
3. find $\bar{n}_k$ and $\bar{r}_{kj}$ using the posterior probability in part two and
4. set the derivatives to 0 and solve.

A Problem!!!

Even though Bock and Aiken simplified the equations, we still need the item parameters to estimate everything else.

Therefore, we must

> *iteratively go through steps 1 – 4 using Newton-Raphson steps and update the item parameter after each step until convergence.*

But ... there is another way.

An Alternative Formulation

The EM algorithm is useful for complex maximum likelihood problems.

EM turns the maximization problem (with unobserved random variables) into a missing data problem.

How does the EM algorithm applies to IRT?

- $\boldsymbol{\theta}$ is unobserved and unobservable.
- $(\mathbf{Y}, \boldsymbol{\theta})$ is the unobserved (complete) data.
- $\mathbf{Y}$ is the observed (incomplete) data.

The Steps

After an initial estimate of item parameters, EM uses two steps:

1. The E Step: compute $E[\log f(\mathbf{Y}, \theta|\mathbf{\Phi})|\mathbf{Y}, \mathbf{\Phi}^p]$
2. The M Step: choose $\mathbf{\Phi}^p$ to maximize the expectation.

Essentially

1. The "E" Step: finds log-likelihood values (**expected** log-likelihood values).
2. The "M" step treats the "E" step output as a genuine log-likelihood.

EM in IRT

The log-likelihood of "observing" the $\mathbf{n}$ and $\mathbf{r}$ vectors is

$$\sum_{k=1}^{q}\sum_{j=1}^{J}\left[r_{kj}\log P_j(X_k) + (n_{kj} - r_{kj})\log Q_j(X_k) + \sum_{k=1}^{q} n_{kj}\log(\alpha_k)\right]$$

Taking expectations with respect to $\mathbf{Y}$ and $\mathbf{\Phi}$ results in

$$\sum_{k=1}^{q}\sum_{j=1}^{J}\left[E(r_{kj}|\mathbf{Y},\,\mathbf{\Phi})\log P_j(X_k) + E[(n_{kj} - r_{kj})|\mathbf{Y},\,\mathbf{\Phi}]\log Q_j(X_k) + \sum_{k=1}^{q} E(n_{kj}|\mathbf{Y},\,\mathbf{\Phi})\log(\alpha_k)\right]$$

And we now have a "posterior likelihood" based on the X_k distribution.

EM in IRT: The Properties

The following are properties of the EM algorithm.

1. $(\mathbf{n}, \mathbf{r})$ is a sufficient statistic for $(\mathbf{Y}, \mathbf{X})$.
2. Maximizing the previous log-likelihood is equivalent to maximizing the E step.
3. The previous log-likelihood has no cross-second derivatives, so the maximization is done item-by-item.
4. Because the two/three parameter logistic IRT models are not exponential family members, the algorithm is not guaranteed to converge.

EM in IRT: The Benefits

The following are benefits of the EM algorithm.

1. Item parameters are consistent for finite length tests (unlike JMLE).
2. The metric of the item parameters is defined by the distribution of examinees.
3. EM imparts a Bayesian-like structure on estimation.

References

- Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science*, *12*, 162–176.
- Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and EM algorithm: A didactic. *Journal of Educational and Behavioral Statistics*, *13*, 243–271.