

2PL模型的EM缺失数据处理方法研究 *

张淑梅¹ 辛 涛² 曾 莉¹ 孙佳楠¹

(¹北京师范大学数学科学学院, 数学与复杂系统教育部重点实验室, 北京, 100875)

(²北京师范大学发展心理研究所, 北京, 100875)

摘 要

项目反应理论(IRT)模型是教育统计与测量中一种十分重要的模型, 它包含项目参数和能力参数. 目前一种常用的估计IRT模型项目参数的方法是由Woodruff和Hanson (1997)应用EM算法给出的, 它用于完全反应数据, 而把能力参数看作缺失数据. 本文将Woodruff的方法推广到处理缺失反应的情况, 基本思想是把能力参数和缺失反应均看作缺失数据, 再运用EM算法估计参数. 通过模拟研究, 在不同被试人数和不同缺失比例的情况下, 本文比较了我们给出的方法和BILOG-MG软件的缺失数据处理方法的参数估计效果. 结果表明, 在大多数情况下, 本文提出的方法能得到更好的估计.

关键词: 项目反应理论, 缺失数据, EM算法.

学科分类号: O212.1.

§1. 引 言

在教育和心理测试中, 经常会遇到被试的项目反应有缺失的情况. 在此情况下如何估计项目反应理论(IRT) (漆书青, 戴海琦, 丁树良, 2002)模型的参数是一项有价值的研究. 目前, 主要有三种IRT模型的缺失数据处理方法. 一种是先删除缺失数据, 然后对完整数据进行估计, Muraki和Bock (1993)便是如此处理缺失. 第二种方法是借补(Little & Rubin, 2002/2004), 首先填充缺失数据, 然后进行参数估计, 如用错误作答填充缺失反应(Ackerman, 1987). 最后一种方法是马尔可夫链蒙特卡洛方法(MCMC; Ross, 2006; Patz & Junker, 1999a; Patz & Junker, 1999b), 它是一种基于模型的方法. 前两种方法都是先把不完全的数据处理成完全的, 然后再对完整数据使用传统的参数估计方法进行估计. 也就是说, 缺失处理不是与参数估计结合在一起的, 而是在参数估计之前. 最后一种方法是基于模型的方法, 它把缺失数据的处理过程同参数估计过程结合了起来. Rubin (2002/2004)指出, 基于模型的方法通常会获得更好的估计, 因此, 本文的研究主要针对基于模型的方法.

对于完整数据, 主要有两种基本的参数估计方法: 极大似然估计和贝叶斯估计(漆书清, 戴海琦, 丁树良, 2002; Baker & Kim, 2004). 在IRT模型的参数估计方法中目前只有一种是基于模型的缺失数据处理方法—MCMC方法, 它是基于贝叶斯理论获得贝叶斯估计的

*本研究受国家自然科学基金(10871026)和教育部新世纪优秀人才支持计划(NCET-07-0097)资助.
本文2010年5月20日收到.

方法, 如Baker (1998), Kim和Cohen (1999), James, Daniel, Allan和Lee (2002). 本文的工作是寻找一种在IRT中可以基于模型处理缺失数据获得参数的极大似然估计的方法.

在IRT中有几种获得参数的极大似然估计的方法. 第一种是联合极大似然估计, 即同时估计能力参数和项目参数. 但是该方法有一个天然的缺陷——项目参数的估计不是相合估计(漆书青, 戴海琦, 丁树良, 2002). 能解决这个问题的一种较好的办法是Bock和Lieberman (Baker & Kim, 2004)给出的边际极大似然估计, 但该方法只在项目数量较少时在计算上可实现. Bock和Aitkin (1981)给出一种实现参数的边际极大似然估计的两步算法. 他们的程序在理论上可接受, 在计算上也可行. Woodruff和Hanson (1997)应用EM算法(Dempster, Laird & Rubin, 1977)于IRT模型获得了参数的极大似然估计, 在他们的程序中, 能力分布被假设为离散型分布. Woodruff和Hanson (1997)已经证明: 如果假设能力分布是已知的, 那么Bock-Aitkin方法就是Woodruff和Hanson的方法的一个特例.

本文主要基于Woodruff和Hanson的想法, 给出了一种在被试的项目反应有缺失的情况下用于估计2PL模型的项目参数的方法.

§2. 两参数逻辑斯蒂克(2PL)模型

本文使用的项目反应模型是广泛使用的2PL模型(漆书青, 戴海琦, 丁树良, 2002). 考虑某一测验中 N 个被试 J 个项目. 记 Y 为 $N \times J$ 的反应为0或1的二分矩阵, $Y_{ij} = 1$ 或0分别表示被试 i 对项目 j 的作答为正确或错误, $1 \leq i \leq N$, $1 \leq j \leq J$. 设被试 i 的能力参数为 θ_i , 第 j 个项目的区分度参数为 α_j , 难度参数为 β_j (α_j, β_j 称为项目参数), 记 $\gamma_j = \alpha_j \times \beta_j$, $\delta_j = (\alpha_j, \gamma_j)$, $j = 1, 2, \dots, J$. 设 $P(\theta_i, \delta_j)$ 表示能力为 θ_i 的被试在项目 j 上正确作答的概率, 则2PL模型为

$$\begin{aligned} P(\theta_i, \delta_j) &= P(y_{ij} = 1 | \theta_i, \alpha_j, \gamma_j) \\ &= \frac{\exp[\alpha_j \theta_i - \gamma_j]}{1 + \exp[\alpha_j \theta_i - \gamma_j]}; \quad j = 1, 2, \dots, J; i = 1, 2, \dots, N. \end{aligned}$$

这里我们增加约束条件 $\theta_i \sim N(0, 1)$, $i = 1, \dots, N$, 即能力参数服从标准正态分布. 可以证明(Tsutakawa, 1982), 该条件与下面这种熟知的约束条件

$$\prod_{j=1}^J \alpha_j = 1, \quad \sum_{j=1}^J \beta_j = 0, \quad (\beta_j = \gamma_j / \alpha_j), \quad \theta_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, N$$

是等价的.

为简化计算, 假设能力参数 θ 取 K 个已知的离散值 q_k , $k = 1, 2, \dots, K$, 对应的未知概率记为 π_k , $k = 1, 2, \dots, K$, 称 $\pi = (\pi_1, \pi_2, \dots, \pi_K)$ 为能力分布参数. 此时, 能力参数的约束为

$$\sum_{k=1}^K q_k \pi_k = 0, \quad \sum_{k=1}^K q_k^2 \pi_k = 1.$$

§3. 方 法

3.1 EM算法

EM算法(Dempster, Laird & Rubin, 1977)是一种为解决含有缺失数据的统计模型中的参数极大似然估计问题的迭代算法. 记完全数据为 $Z = (Z_{\text{obs}}, Z_{\text{mis}})$, 其中 Z_{obs} 表示数据的观测部分, Z_{mis} 表示数据的缺失部分. 假设 Z 有密度 $f(Z|\Delta)$, 这里 Δ 为未知参数. 为了极大化观测数据的似然函数 $L(\Delta|Z_{\text{obs}}) = f(Z_{\text{obs}}|\Delta)$, 该算法利用了完全数据似然函数 $L(\Delta|Z_{\text{obs}}, Z_{\text{mis}}) = L(\Delta|Z) = f(Z|\Delta)$. 如前所述, EM算法是一种迭代算法, 每次迭代包括两个步骤: E (求期望)步, M (最大化)步. 具体地说, 在迭代的第 s 步时, E步要计算如下的条件期望

$$\begin{aligned} Q(\Delta|\Delta^{(s)}) &= E_{Z_{\text{mis}}|Z_{\text{obs}}, \Delta=\Delta^{(s)}} \log L(\Delta|Z_{\text{obs}}, Z_{\text{mis}}) \\ &= \int \log L(\Delta|Z) f(Z_{\text{mis}}|Z_{\text{obs}}, \Delta = \Delta^{(s)}) dZ_{\text{mis}}. \end{aligned} \quad (3.1)$$

(3.1)式中的 $E_{Z_{\text{mis}}|Z_{\text{obs}}, \Delta=\Delta^{(s)}}[\cdot]$ 是在给定现有参数 $\Delta^{(s)}$ 和观测数据 Z_{obs} 的条件下, 基于缺失数据 Z_{mis} 的条件分布对完全数据似然函数 $\log L(\Delta|Z_{\text{obs}}, Z_{\text{mis}})$ 求期望, 因此(3.1)式不含有未知变量 Z_{mis} , 含有的未知参数只有 Δ .

M步是: 将E步获得的包含未知参数 Δ 的完全数据似然函数的期望进行极大化, 从而确定参数值 $\Delta^{(s+1)}$. 再将M步得到的未知参数估计值带入E步……如此迭代, 直至收敛.

3.2 完全反应下的EM算法

这一节将简单介绍Woodruff和Hanson (1997)的针对项目完全反应获得2PL模型参数的极大似然估计的EM算法. 将EM算法应用到IRT的参数估计的基本思想是: 不把被试的能力 $\theta = (\theta_1, \theta_2, \dots, \theta_N)$ 看作参数, 而看作缺失数据 Z_{mis} . 此时观测数据 Z_{obs} 为反应矩阵 Y , 缺失数据为被试的能力 θ , 完全数据为 $Z = (Y, \theta)$. 参数包括项目参数 $\Delta = [\delta_1, \dots, \delta_J]$ 和能力分布参数 $\pi = (\pi_1, \pi_2, \dots, \pi_K)$. 记 $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iJ})$ 为被试 i 的作答, 则 $Y = (\mathbf{y}_1^T, \dots, \mathbf{y}_N^T)^T$ (其中 \mathbf{y}_i^T 表示向量 \mathbf{y}_i 的转置). 根据(3.1)式, E步中计算的条件期望推导如下

$$\begin{aligned} Q(\Delta, \pi|\Delta^{(s)}, \pi^{(s)}) &= E_{Z_{\text{mis}}|Z_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [\log L(\Delta, \pi|Z_{\text{obs}}, Z_{\text{mis}})] \\ &= E_{\theta|Y, \Delta^{(s)}, \pi^{(s)}} [\log L(\Delta, \pi|Y, \theta)] \\ &= E_{\theta|Y, \Delta^{(s)}, \pi^{(s)}} \left[\log \prod_{i=1}^N f(\mathbf{y}_i, \theta_i|\Delta, \pi) \right] \\ &= \sum_{i=1}^N E_{\theta_i|\mathbf{y}_i, \Delta^{(s)}, \pi^{(s)}} [\log f(\mathbf{y}_i, \theta_i|\Delta, \pi)]. \end{aligned} \quad (3.2)$$

通过上述推导, Woodruff和Hanson已经证明条件期望 $Q(\Delta, \pi|\Delta^{(s)}, \pi^{(s)})$ 可以写成 $\phi(\Delta)$ 、 $\psi(\pi)$ 两项的和, 即

$$Q(\Delta, \pi|\Delta^{(s)}, \pi^{(s)}) = \phi(\Delta) + \psi(\pi), \quad (3.3)$$

其中(3.3)式等号右端的第一项含有的未知参数只有项目参数, 而第二项含有的未知参数只有能力分布参数. 具体表达如下

$$\phi(\Delta) = \sum_{k=1}^K \sum_{j=1}^J \{ \log[P(q_k, \delta_j)] r_{jk}^{(s)} + \log[Q(q_k, \delta_j)] (n_k^{(s)} - r_{jk}^{(s)}) \}, \quad (3.4)$$

$$\psi(\pi) = \sum_{k=1}^K \log(\pi_k) n_k^{(s)}, \quad (3.5)$$

其中

$$n_k^{(s)} = \sum_{i=1}^N \frac{f(\mathbf{y}_i|q_k, \Delta^{(s)}) \pi_k^{(s)}}{\sum_{k'=1}^K f(\mathbf{y}_i|q_{k'}, \Delta^{(s)}) \pi_{k'}^{(s)}}, \quad (3.6)$$

$$r_{jk}^{(s)} = \sum_{i=1}^N \frac{y_{ij} f(\mathbf{y}_i|q_k, \Delta^{(s)}) \pi_k^{(s)}}{\sum_{k'=1}^K f(\mathbf{y}_i|q_{k'}, \Delta^{(s)}) \pi_{k'}^{(s)}}, \quad (3.7)$$

$$f(\mathbf{y}_i|q_k, \Delta^{(s)}) = \prod_{j=1}^J P(q_k, \delta_j^{(s)})^{y_{ij}} [1 - P(q_k, \delta_j^{(s)})]^{1-y_{ij}}. \quad (3.8)$$

在(3.4)式中, 记 $Q(q_k, \delta_j) \triangleq 1 - P(q_k, \delta_j)$, 与 $\log[Q(q_k, \delta_j)]$ 均含有未知的项目参数 δ_j . (3.6)式的 $n_k^{(s)}$ 与(3.7)式的 $r_{jk}^{(s)}$ 称为人工数据, 二者均不含任何未知参数. $n_k^{(s)}$ 可理解为在 N 个被试中能力水平为 q_k 的被试数目的期望, $r_{jk}^{(s)}$ 可理解为在 N 个被试中具有能力水平 q_k 的被试答对第 j 个项目的期望. 应用EM算法的第 s 步迭代中, E步: 利用第 $s-1$ 步得到的参数估计值 $\Delta^{(s)}$ 和 $\pi_k^{(s)}$ 计算 $n_k^{(s)}$ 和 $r_{jk}^{(s)}$ (首次迭代利用的初值记为 $\Delta^{(0)}$ 和 $\pi_k^{(0)}$).

M步: 将E步计算出的 $n_k^{(s)}$ 和 $r_{jk}^{(s)}$ 带入 $\phi(\Delta)$ 和 $\psi(\pi)$, 对两项分别极大化可得参数 Δ 和 π 的估计值 $\Delta^{(s+1)}$ 和 $\pi_k^{(s+1)}$. Woodruff和Hanson在极大化 $\phi(\Delta)$ 时使用的是Newton-Raphson算法, 而极大化 $\psi(\pi)$ 时参数的解可直接求得, 为

$$\pi_k^{(s+1)} = \frac{n_k^{(s)}}{\sum_{k'=1}^K n_{k'}^{(s)}}. \quad (3.9)$$

再把M步得到的参数估计值带入E步..... 如此迭代, 直至收敛.

3.3 缺失反应下的EM算法

在实际测验中, 被试的作答常出现缺失. 我们将上节Woodruff和Hanson(1997)的思想应用于项目反应有缺失的情况. 假设反应矩阵 Y 有缺失, 那么可以把反应矩阵写成两

个部分: $Y = (Y_{\text{obs}}, Y_{\text{mis}})$. 其中 Y_{obs} 代表反应矩阵的观测部分, Y_{mis} 表示反应矩阵的缺失部分. 所以完全数据可写为 $Z = [Y, \theta] = [Y_{\text{obs}}, Y_{\text{mis}}, \theta] = [Y_{\text{obs}}, (Y_{\text{mis}}, \theta)]$, 此时观测数据为 $Z_{\text{obs}} = Y_{\text{obs}}$, 缺失数据为 $Z_{\text{mis}} = (Y_{\text{mis}}, \theta)$. 现在, 我们对这种数据应用EM算法. 根据(3.1)式, E步中计算的条件期望的推导为

$$\begin{aligned}
 & Q(\Delta, \pi | \Delta^{(s)}, \pi^{(s)}) \\
 &= E_{Z_{\text{mis}} | Z_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [l(\Delta, \pi | Z_{\text{obs}}, Z_{\text{mis}})] \\
 &= E_{(Y_{\text{mis}}, \theta) | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [l(\Delta, \pi | Y, \theta)] \\
 &= E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} E_{\theta | Y_{\text{obs}}, Y_{\text{mis}}, \Delta^{(s)}, \pi^{(s)}} [l(\Delta, \pi | Y, \theta)] \\
 &= E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [\phi(\Delta; Y_{\text{mis}}) + \psi(\pi; Y_{\text{mis}})] \\
 &= E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [\phi(\Delta; Y_{\text{mis}})] + E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [\psi(\pi; Y_{\text{mis}})] \\
 &\triangleq \phi^*(\Delta) + \psi^*(\pi), \tag{3.10}
 \end{aligned}$$

其中

$$\phi^*(\Delta) \triangleq E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [\phi(\Delta; Y_{\text{mis}})], \psi^*(\pi) \triangleq E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [\psi(\pi; Y_{\text{mis}})].$$

将(3.10)式同项目完全反应下的(3.3)式作比较可发现: 二者的相似之处是都可以写成两项的和, 第一项包含的未知参数只有项目参数 Δ , 第二项包含的未知参数只有能力分布参数 π ; 二者都是在给定第 $s-1$ 步得到的项目参数估计值 $\Delta^{(s)}$ 和能力分布参数估计值 $\pi^{(s)}$ 的条件下求条件期望. 二者的不同之处是: (3.10)式既关于能力参数 θ 求条件期望, 又关于缺失反应 Y_{mis} 求条件期望; 而(3.3)式由于是针对完全反应项目的情形, 只包含对能力参数 θ 求条件期望.

为进一步说明, 我们引入一些符号. Y_{obs} 中对应第 i 个被试的观测记为 $Y_{\text{obs}, i}$, Y_{mis} 中对应第 i 个被试的缺失反应记为 $Y_{\text{mis}, i}$, 将 $y_{\text{mis}, i}$ 的长度记为 n_i , 即被试 i 的缺失项目数是 n_i . 记 $\mathbf{y}_i = (y_{\text{obs}, i}, y_{\text{mis}, i})$. 如果被试 i 回答了项目 j , 记 $M_{ij} = 0$; 否则, 记 $M_{ij} = 1$.

基于Woodruff和Hanson的思想, 我们通过对项目反应有缺失的情况进行推导, 得出

$$\begin{aligned}
 \phi^*(\Delta) &= E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} [\phi(\Delta; Y_{\text{mis}})] \\
 &= E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} \sum_{k=1}^K \sum_{j=1}^J \{ \log[P(q_k, \delta_j)] r_{jk}^{(s)} + \log[Q(q_k, \delta_j)] (n_k^{(s)} - r_{jk}^{(s)}) \} \\
 &= \sum_{k=1}^K \sum_{j=1}^J \{ \log[P(q_k, \delta_j)] E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} (r_{jk}^{(s)}) \\
 &\quad + \log[1 - P(q_k, \delta_j)] [E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} (n_k^{(s)}) - E_{Y_{\text{mis}} | Y_{\text{obs}}, \Delta^{(s)}, \pi^{(s)}} (r_{jk}^{(s)})] \},
 \end{aligned}$$

其中

$$\begin{aligned}
 & E_{Y_{\text{mis}}|Y_{\text{obs}},\Delta^{(s)},\pi^{(s)}}(n_k^{(s)}) \\
 &= \sum_{i=1}^N E_{y_{\text{mis},i}|y_{\text{obs},i},\Delta^{(s)},\pi^{(s)}} \sum_{k'=1}^K \frac{f(y_i|q_k, \Delta^{(s)})\pi_k^{(s)}}{\sum_{k'=1}^K f(y_i|q_{k'}, \Delta^{(s)})\pi_{k'}^{(s)}} \\
 &= \sum_{i=1}^N \sum_{\substack{y_{\text{mis},i}=(I_1,\dots,I_{n_i}); \\ I_t=0,1;t=1,\dots,n_i}} \sum_{k'=1}^K \frac{f(y_i|q_k, \Delta^{(s)})\pi_k^{(s)}}{\sum_{k'=1}^K f(y_i|q_{k'}, \Delta^{(s)})\pi_{k'}^{(s)}} \Big|_{y_{\text{mis},i}=(I_1,\dots,I_{n_i})} \\
 &\quad \times P[y_{\text{mis},i}=(I_1,\dots,I_{n_i})|\Delta^{(s)},\pi^{(s)}], \\
 & E_{Y_{\text{mis}}|Y_{\text{obs}},\Delta^{(s)},\pi^{(s)}}(r_{jk}^{(s)}) \\
 &= \sum_{i=1}^N E_{y_{\text{mis},i}|y_{\text{obs},i},\Delta^{(s)},\pi^{(s)}} \sum_{k'=1}^K \frac{y_{ij}f(y_i|q_k, \Delta^{(s)})\pi_k^{(s)}}{\sum_{k'=1}^K f(y_i|q_{k'}, \Delta^{(s)})\pi_{k'}^{(s)}} \\
 &= \sum_{i=1}^N \sum_{\substack{y_{\text{mis},i}=(I_1,\dots,I_{n_i}); \\ I_t=0,1;t=1,\dots,n_i}} \sum_{k'=1}^K \frac{y_{ij}f(y_i|q_k, \Delta^{(s)})\pi_k^{(s)}}{\sum_{k'=1}^K f(y_i|q_{k'}, \Delta^{(s)})\pi_{k'}^{(s)}} \Big|_{y_{\text{mis},i}=(I_1,\dots,I_{n_i})} \\
 &\quad \times P[y_{\text{mis},i}=(I_1,\dots,I_{n_i})|\Delta^{(s)},\pi^{(s)}], \\
 & P[y_{\text{mis},i}=(I_1,\dots,I_{n_i})|\Delta^{(s)},\pi^{(s)}] \\
 &= \sum_{m=1}^K P(y_{\text{mis},i}=(I_1,\dots,I_{n_i})|\theta=q_m, \Delta^{(s)}, \pi^{(s)})P(\theta=q_m|\Delta^{(s)}, \pi^{(s)}) \\
 &= \sum_{m=1}^K \pi_m^{(s)} \prod_{t': M_{it'}=1} [P(q_m, \delta_{t'}^{(s)})^{y_{it'}} Q(q_m, \delta_{t'}^{(s)})^{1-y_{it'}}].
 \end{aligned}$$

由于人工数据中一些项目的反应含有缺失, 所以本节的上述表达式与3.2节的不同之处是本节利用了条件期望将缺失 Y_{mis} 从人工数据中积分掉; 其中, 条件期望对应的条件概率是在完全随机缺失(MCAR)的机制下求出的. 做此处理后, 本节的人工数据的计算结果不含有任何未知量. 所以, 当被试的项目反应存在缺失时, 应用EM算法的步骤中, E步: 利用第 $s-1$ 步得出的参数估计值 $\Delta^{(s)}$ 和 $\pi_k^{(s)}$, 计算 $E_{Y_{\text{mis}}|Y_{\text{obs}},\Delta^{(s)},\pi^{(s)}}(n_k^{(s)})$ 和 $E_{Y_{\text{mis}}|Y_{\text{obs}},\Delta^{(s)},\pi^{(s)}}(r_{jk}^{(s)})$ (首次迭代利用的初值记为 $\Delta^{(0)}$ 和 $\pi_k^{(0)}$).

M步: 将E步计算出的 $E_{Y_{\text{mis}}|Y_{\text{obs}},\Delta^{(s)},\pi^{(s)}}(n_k^{(s)})$ 和 $E_{Y_{\text{mis}}|Y_{\text{obs}},\Delta^{(s)},\pi^{(s)}}(r_{jk}^{(s)})$ 带入 $\phi^*(\Delta)$ 和 $\psi^*(\pi)$ 中, 对二者分别极大化可以得到参数 Δ 和 π 的估计 $\Delta^{(s+1)}$ 和 $\pi_k^{(s+1)}$. 在 $\phi^*(\Delta)$ 极大化中我们同样使用Newton-Raphson算法. 而对于 $\psi^*(\pi)$ 的极大化, 类似于Woodruff和Hanson的推导, 求得的能力分布参数 π 的解为

$$\pi_k^{(s+1)} = \frac{E_{Y_{\text{mis}}|Y_{\text{obs}},\Delta^{(s)},\pi^{(s)}}(n_k^{(s)})}{\sum_{k'=1}^K E_{Y_{\text{mis}}|Y_{\text{obs}},\Delta^{(s)},\pi^{(s)}}(n_{k'}^{(s)})}. \quad (3.11)$$

再把M步得到的参数估计值带入E步..... 如此迭代, 直至收敛.

可以看出(3.11)式同(3.9)式不同, (3.11)式包含了对人工数据 $n_k^{(s)}$ 关于项目的缺失反应 Y_{mis} 求条件期望的步骤.

§4. 模拟研究

当被试的反应矩阵有缺失的时候, 常用软件BILOG-MG默认的处理方法是把缺失反应插补为0, 即把缺失反应处理成错误作答. 下面采用模拟的方法, 把我们的缺失数据处理方法同BILOG-MG软件的处理方法进行比较.

考虑到参数估计可能受到被试人数 N , 缺失比例 p 等因素的影响, 在模拟被试的反应矩阵时我们考虑了四种被试人数($N = 2000, 1000, 500, 100$), 三种缺失比例($p = 0.05, 0.1, 0.25$). 因此一共有12种组合情况.

取 $\alpha \sim \log \text{norm}(0, 0.5)$, $\gamma \sim N(0, 2)$, $\theta \sim N(0, 1)$. 根据各个参数的分布分别模拟产生出参数真值, 带入2PL模型计算得到每个被试在每个项目上的正确作答概率, 再根据均匀分布模拟出判定每个项目的反应正确与否的临界值, 将被试的正确作答概率同其进行比较, 大于临界值则记为正确作答, 否则记为错误作答, 即得到反应矩阵. 接着模拟缺失, 对每个被试每个项目产生均匀分布的随机数, 若该随机数小于缺失比例 p , 则认为该被试在这个项目上缺失. 从而得到含有缺失反应的反应矩阵. 此时我们假设缺失机制是完全随机缺失(MCAR; Little & Rubin, 2002/2004).

对于 N, p 的12种组合情况分别模拟含有缺失的反应矩阵, 再分别运用我们的方法和BILOG软件估计项目参数, 并进行比较分析. 这里, 每种组合情况均重复模拟20次.

为衡量两种方法揭示真实参数的性能, 运用估计的均方误差RMSE (root mean square errors), 偏差BIAS和标准误SE (standard errors), 公式如下

$$\begin{aligned} \text{RMSE}(\hat{\alpha}) &= \sqrt{\frac{\sum_m^M \sum_j^J (\hat{\alpha}_{mj} - \alpha_j)^2}{MJ}}, & \text{RMSE}(\hat{\gamma}) &= \sqrt{\frac{\sum_m^M \sum_j^J (\hat{\gamma}_{mj} - \gamma_j)^2}{MJ}}, \\ \text{BIAS}(\hat{\alpha}) &= \frac{\sum_{j=1}^J \sum_{m=1}^M (\hat{\alpha}_{mj} - \alpha_j)}{MJ}, & \text{BIAS}(\hat{\gamma}) &= \frac{\sum_{j=1}^J \sum_{m=1}^M (\hat{\gamma}_{mj} - \gamma_j)}{MJ}, \\ \text{SE}(\hat{\alpha}) &= \sqrt{\frac{\sum_{j=1}^J \sum_{m=1}^M \left(\hat{\alpha}_{mj} - \frac{\sum_{m=1}^M \hat{\alpha}_{mj}}{M} \right)^2}{MJ}}, & \text{SE}(\hat{\gamma}) &= \sqrt{\frac{\sum_{j=1}^J \sum_{m=1}^M \left(\hat{\gamma}_{mj} - \frac{\sum_{m=1}^M \hat{\gamma}_{mj}}{M} \right)^2}{MJ}}, \end{aligned}$$

其中 M 为模拟重复次数, 这里 $M = 20$.

§5. 研究结果

为叙述方便, 下面均以“EM算法”指代3.3节用于处理项目反应缺失的EM算法.

为了检测我们自编的程序, 首先在项目反应无缺失时比较了EM算法和BILOG-MG软件的参数估计结果(见表1). 结果表明: EM算法对斜率参数的估计较BILOG而言略差一点, 但截距参数的估计较BILOG而言略好些, 这说明两种方法在参数估计上的精度大致相当; 而产生这种差别的原因可能是EM算法同BILOG在程序的内部优化处理上存在着差异.

表1 项目反应无缺失的EM与BILOG的RMSE结果比较(保留2位小数)

被试人数	参数类型	EM	BILOG
1000	斜率参数	0.17	0.15
1000	截距参数	0.17	0.18

下面着重比较反应矩阵有缺失时, BILOG软件和EM算法在项目参数估计精度上的差异. 表2是这些关于斜率和截距参数估计的RMSE结果. 表3是两种方法的评价指标RMSE之差. 表4是两种方法关于斜率和截距参数估计的BIAS. 表5是两种方法关于斜率和截距参数估计的SE.

表2 项目反应有缺失的EM与BILOG的RMSE结果比较(保留2位小数)

		斜率参数		截距参数	
缺失比例	被试人数	EM	BILOG	EM	BILOG
0.05	2000	0.15	0.19	0.13	0.26
	1000	0.18	0.22	0.14	0.26
	500	0.22	0.24	0.19	0.28
	100	0.73	0.43	0.71	0.49
0.1	2000	0.26	0.29	0.18	0.41
	1000	0.27	0.30	0.16	0.41
	500	0.28	0.32	0.19	0.43
	100	0.58	0.48	0.62	0.56
0.25	2000	0.51	0.48	0.29	0.80
	1000	0.52	0.48	0.25	0.81
	500	0.51	0.47	0.27	0.81
	100	0.61	0.59	0.52	0.87

表3 项目反应有缺失的EM与BILOG的RMSE差值的结果比较(保留2位小数)

缺失比例	被试人数	EM-BILOG (α)	EM-BILOG (γ)
0.05	2000	-0.04	-0.12
	1000	-0.04	-0.11
	500	-0.01	-0.09
	100	0.30	0.21
0.1	2000	-0.03	-0.23
	1000	-0.04	-0.25
	500	-0.03	-0.24
	100	0.10	0.07
0.25	2000	0.03	-0.52
	1000	0.03	-0.56
	500	0.03	-0.54
	100	0.02	-0.35

表4 项目反应有缺失的EM与BILOG的BIAS结果比较(保留2位小数)

		斜率参数		截距参数	
缺失比例	被试人数	EM	BILOG	EM	BILOG
0.05	2000	-0.09	-0.11	-0.03	0.18
	1000	-0.08	-0.10	-0.01	0.21
	500	-0.07	-0.11	0.00	0.18
	100	0.08	-0.13	-0.01	0.11
0.1	2000	-0.20	-0.20	-0.05	0.33
	1000	-0.19	-0.19	-0.03	0.35
	500	-0.19	-0.20	-0.03	0.33
	100	-0.08	-0.20	-0.03	0.26
0.25	2000	-0.44	-0.38	-0.09	0.69
	1000	-0.43	-0.37	-0.07	0.71
	500	-0.43	-0.37	-0.08	0.70
	100	-0.38	-0.37	-0.10	0.64

表5 项目反应有缺失的EM与BILOG的SE结果比较(保留2位小数)

		斜率参数		截距参数	
缺失比例	被试人数	EM	BILOG	EM	BILOG
0.05	2000	0.09	0.10	0.07	0.07
	1000	0.12	0.12	0.09	0.09
	500	0.17	0.15	0.13	0.12
	100	0.57	0.26	0.40	0.26
0.1	2000	0.08	0.10	0.06	0.08
	1000	0.10	0.11	0.08	0.09
	500	0.15	0.15	0.12	0.13
	100	0.44	0.26	0.36	0.26
0.25	2000	0.07	0.10	0.06	0.08
	1000	0.09	0.11	0.08	0.08
	500	0.12	0.14	0.11	0.12
	100	0.32	0.23	0.30	0.26

5.1 两种方法下参数估计RMSE的比较

从表2和表3的斜率参数的对应结果可以看出: 样本量为2000、1000、500时, 综合比较后发现, 我们运用EM算法处理缺失的方法对应的两类参数的估计精度优于BILOG-MG将缺失用0借补(看作错误作答)的方法. 其中, 两种方法对应斜率参数的RMSE相差不大, 但对截距参数来说, 我们的缺失处理方法的RMSE小于BILOG-MG的借补法对应的RMSE, 且十分明显; 这种差别也可从表3的BIAS结果比较中看出. 当样本量较少($N = 100$)时, 两种方法均表现较差, 尤其是EM算法得到的斜率参数和BILOG-MG得到的截距参数; 这与样本量较少时两种方法的参数估计均不稳定有关.

5.1.1 比较被试人数对参数估计RMSE的影响

首先, 我们想要观察被试人数对两种方法的表现的效应. 为了更直观地比较, 我们画出图1, 它是12种组合情况的两种方法的RMSE散点图. 图1的左边是斜率参数的RMSE, 右边是截距参数的RMSE. 图中三角形符号代表EM算法, 加号代表BILOG-MG.

为了比较两种方法的参数估计效果受被试人数的影响, 我们首先固定了缺失比例. 每张图中左边的一组缺失比例为0.05, 中间一组缺失比例为0.1, 右边一组的缺失比例为0.25. 我们可以在每个组内观察被试人数的效应. 在表3中, 在0.05的缺失比例下, 斜率参数估计

的两种方法的RMSE之差的范围是-0.04到0.30; 当缺失比例为0.1时, 两种方法的RMSE之差的范围是-0.04到0.10; 当缺失比例为0.25时, 两种方法的RMSE之差的范围是0.02到0.03. 对截距参数来说, 当缺失比例为0.05时, 两种方法的RMSE之差的范围是-0.12到0.21; 当缺失比例为0.1时, 两种方法的RMSE之差的范围是-0.25到0.07; 当缺失比例为0.25时, 两种方法的RMSE之差的范围是-0.52到-0.35. 因此, 就总体而言, 两种方法对于斜率参数的估计精度较为类似; 而对于截距参数而言, 两种方法的差异非常明显. 除被试人数 $N = 100$ 之外, 在缺失比例较小的条件下, 两种方法的差异相对较小; 而在缺失比例较大的条件下, 两种方法的差异变得更大.

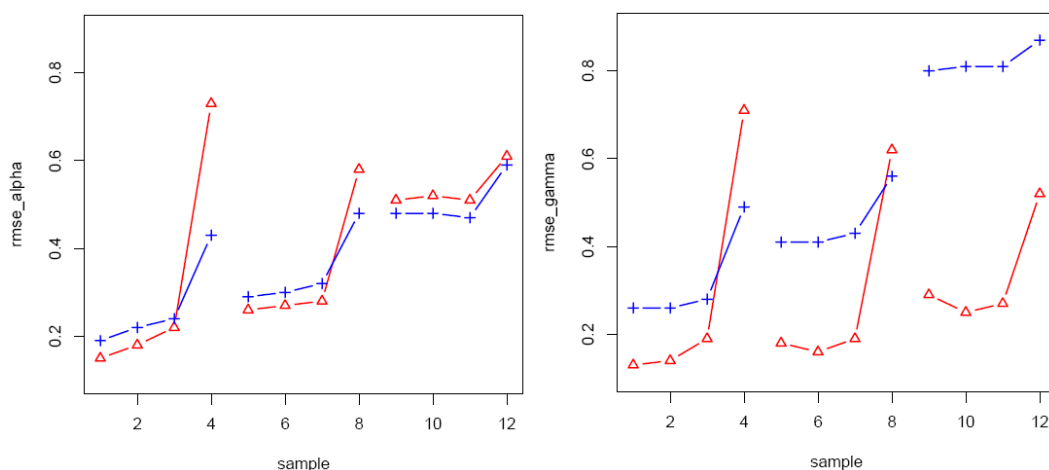


图1 项目反应有缺失的EM与BILOG的RMSE结果对比(比较样本量的效应)

由图1可以看出: 随着被试人数从较多(2000个被试)到中等数目(1000、500个被试), 两种方法的RMSE都变化较小. 但是随着被试人数从中等数目(500个被试)到较少(100个被试), 两种方法的估计效果很快变差. 两种方法在小样本的情况下都表现得不是很好.

5.1.2 比较缺失比例对参数估计RMSE的影响

下面, 我们观察两种方法的表现受缺失比例的影响, 因此, 我们需要固定样本量. 图2给出了12种组合情况的两种方法的RMSE散点图. 每张图中四组从左至右依次对应样本容量为2000, 1000, 500, 100.

在图2中, 对斜率参数来说, 当样本量为2000时, 两种方法的RMSE之差的范围是-0.04到0.03; 当样本量为1000时, 两种方法的RMSE之差的范围是-0.04到0.03; 当样本量为500时, 两种方法的RMSE之差的范围是-0.03到0.03; 当样本量为100时, 两种方法的RMSE之差的范围是-0.03到0.30. 对于截距参数, 当样本量为2000时, 两种方法的RMSE之差的范围是-0.52到-0.12; 当样本量为1000时, 两种方法的RMSE之差的范围是-0.56到-0.11; 当样本

量为500时, 两种方法的RMSE之差的范围是-0.54到-0.09; 当样本量为100时, 两种方法的RMSE之差的范围是-0.35到0.21. 因此, 就总体而言, 两种方法对于斜率参数的估计精度较为类似, 尤其在缺失比例较小的时候. 而对于截距参数而言, 两种方法的差异则非常明显. 无论较大样本($N = 2000$)还是中等样本($N = 1000, 500$)的条件下, 两种方法的差异都比较稳定; 而在样本量较小($N = 100$)的条件下, 两种方法的差异变得更大, 且相对其他样本量而言更不稳定.

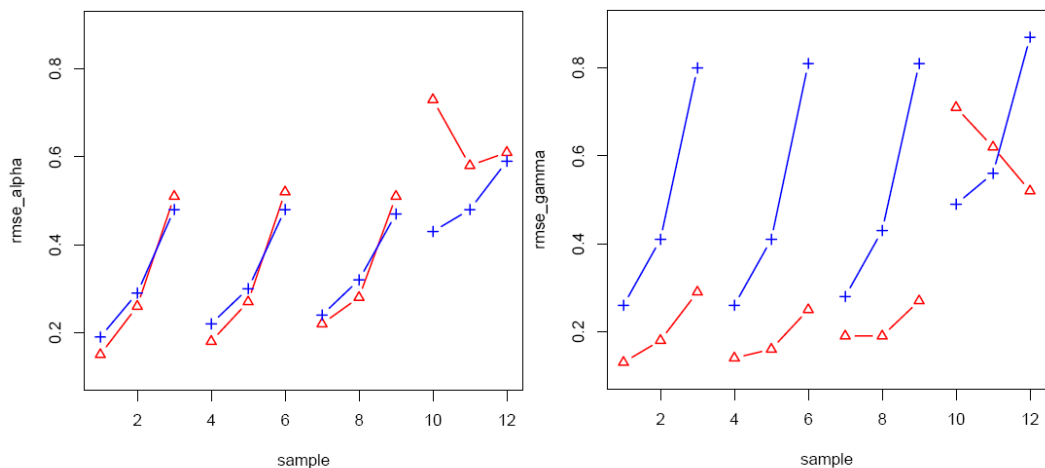


图2 项目反应有缺失的EM与BILOG的RMSE结果比较(比较缺失比例的效应)

从图2可以看出: 在每组内, 随着缺失比例的增加, 两种方法的参数估计均逐渐变差. 对斜率参数而言, 当给定的样本量较大($N = 2000$)和中等($N = 1000, 500$)的时候, EM算法比BILOG稍好.

5.2 两种方法下参数估计BIAS的比较

由表4, 比较斜率参数的估计, 两种缺失处理方法的偏差相差不大. 缺失比例为0.05时, 我们运用EM算法处理缺失的方法有略小的偏差; 缺失比例为0.10时, 两种方法的偏差几乎相同; 缺失比例为0.25时, BILOG-MG的借补法对应的偏差略小些. 比较截距参数的估计, 我们运用EM算法处理缺失数据的方法明显地比BILOG-MG具有更小的偏差! 随着缺失比例由小到大, 可以发现二者偏差的差距逐渐变大! 即在2PL模型下, 缺失数据的确导致了参数估计的系统误差的增大, 其中BILOG-MG对截距参数的估计的系统误差增大得更为明显! 另外, 对相同缺失比例而不同被试人数的情形, 除 $N = 100$ 之外, 两种估计方法的偏差较稳定, 并未明显受到被试人数的影响.

5.3 两种方法下参数估计SE的比较

由表5, 可知: 对两类参数, 不论哪种缺失比例下, 样本容量为 $N = 2000$ 和 $N = 1000$ 时, 均表明我们使用EM算法比BILOG-MG的借补法具有略小的随机误差(SE); 样本容量为 $N = 500$ 时, 两种方法相差不大. 但 $N = 100$ 时, 我们使用的EM算法参数估计的随机误差较大. 另外, 在2PL模型下, 参数估计的随机误差并未随着缺失比例的增加而增加.

§6. 讨 论

贝叶斯估计和极大似然估计是IRT中处理完全数据的两种参数估计方法. Rubin (2002/2004)指出, 基于模型的方法通常比其他的缺失数据处理方法更好. 目前, 只有一种基于模型的缺失数据处理方法应用于IRT中——MCMC方法, MCMC方法是基于贝叶斯理论的. 目前还没有以极大似然估计为基础的基于模型的方法, 本文探索的正是以极大似然估计为基础的基于模型的缺失数据处理方法. 我们完成了这种方法的理论推导并用R软件实现了算法编程. 通过模拟研究, 综合比较EM算法与BILOG-MG 3.0软件的缺失数据处理方法下的参数估计结果, 可以看出:

一方面, 当样本容量为2000、1000、500时, 对三种不同的缺失比例, 大多数情况下, 我们运用EM算法处理缺失获得参数估计值的方法, 比BILOG-MG先借补再估计参数的方法具有更小的均方误差和偏差, 尤其对截距参数而言更为明显; 同时, 两种方法具有相近的参数估计的标准误. 也就是说, 样本容量较大时, 两种缺失处理方法得出的参数估计值的标准误差异不大, 但从估计的均方误差和偏差角度看, 我们的方法比BILOG-MG的参数估计精度会更高. 在实际应用中, 判断样本容量的大小应参照测验包含的项目数. 我们认为被试数目至少达到测验项目数的50倍, 可以看作较大的样本量.

BILOG在某些情况下估计效果不佳的原因是它把所有的缺失反应处理成错误作答, 但事实上, 在某些情况下学生没有完成所有的项目可能是因为时间不够, 或者他们没有看到这个项目, 或还有其他的原因. 把所有的缺失反应处理成错误作答, 可能会导致参数估计的偏差, 本文的模拟研究便证实了这一点. 相比之下, EM算法采用了把缺失处理和参数估计结合在一起的方法, 不需要先填充缺失数据.

另一方面, 我们注意到: 当样本量为100时, 不论哪种缺失比例下, 对15个题目的测验而言, 使用两种缺失处理方法都不足以得到斜率、截距参数较稳定的估计结果. 而且, 当样本容量 $N = 100$ 时, 相比BILOG-MG, 我们应用EM算法处理缺失的方法对斜率参数和截距参数的估计具有较大的随机误差. 因此, 在小样本容量下, 不建议采用我们的方法.

另外, 虽然EM算法在大多数情形下, 可以得到更好的参数估计, 但在某些情况下这种方法也有一定的缺点——当缺失比例高时, 这种方法会花费很多时间. 从方法的推导过程看, 它需要计算每个人的所有可能的缺失反应模式. 当某个人的缺失项目数很大时, 这种缺失

的可能模式的数量将会很多, 因此它比较耗时. 但是, 牺牲时间的同时可以获得更精确的估计结果.

本文的模拟研究基于完全随机缺失 (MCAR; Little & Rubin, 2002/2004) 的缺失机制. 进一步研究可基于随机缺失 (MAR; Little & Rubin, 2002/2004) 或非随机缺失 (MNAR; Little & Rubin, 2002/2004).

参 考 文 献

- [1] Baker, F.B. and Kim, S.-H., *Item Response Theory: Parameter Estimation Techniques* (2nd ed.), New York: Marcell Dekker, 2004.
- [2] Baker, F.B., An investigation of the item parameter recovery characteristics of a Gibbs sampling approach, *Applied Psychological Measurement*, **22**(1998), 153–169.
- [3] Bock, R.D. and Aitken, M., Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, *Psychometrika*, **46**(1981), 443–459.
- [4] Dempster, A.P., Laird, N.M. and Rubin, D.B., Maximum likelihood for incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Ser. B*, **39**(1977), 1–38.
- [5] James, A.W., Daniel, M.B., Allan, S.C. and Lee, Y.-S., Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation, *Applied Psychological Measurement*, **26**(2002), 339.
- [6] Kim, S.-H. and Cohen, A.S., Accuracy of parameter estimation in Gibbs sampling under the two-parameter logistic model, *Paper Presented at the Annual Meeting of the American Educational Research Association*, Montreal, Canada, 1999.4.
- [7] Little, R.J.A. and Rubin, D.B., *Statistical Analysis with Missing Data* (2nd ed.), (Sun, S.Z., Trans.), New York: John Wiley and Sons, 2004 (Original work published 2002).
- [8] Muraki, E. and Bock, R.D., *PARSCALE: IRT Based Test Scoring and Item Analysis for Graded Open-Ended Exercises and Performance Tasks*, Chicago: Scientific software Int, 1993.
- [9] Patz, R.J. and Junker, B.W., A straightforward approach to Markov chain Monte Carlo methods for item response models, *Journal of Educational and Behavioral Statistics*, **24**(1999a), 146–178.
- [10] Patz, R.J. and Junker, B.W., Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses, *Journal of Educational and Behavioral Statistics*, **24**(1999b), 342–366.
- [11] 漆书清, 戴海琦, 丁树良, 现代教育与心理测量学原理, 北京, 高等教育出版社, 2002.
- [12] Sheldon, M.R., *Simulation* (4th ed.), Beijing, China, Post and Telecom Press, 2006.
- [13] Tsutakawa, R.K., Estimation of item parameters and the EGM algorithm, In David J. Weiss (Eds.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference*, 1985, 180–188.
- [14] Woodruff, D. and Hanson B.A., Estimation for item response models using the EM algorithm for finite mixtures, *Presented at the Annual Meeting of the Psychometric Society*, Tennessee, Gatlinburg, 1997.6.

Extension of EM Algorithm for Finite Mixture in IRT for Missing Response Data

ZHANG SHUMEI¹ XIN TAO² ZENG LI¹ SUN JIANAN¹

(¹*School of Mathematical Sciences, Beijing Normal University,*

Laboratory of Mathematics and Complex Systems, Beijing, 100875)

(²*Institute of Developmental Psychology, Beijing Normal University, Beijing, 100875)*

Item Response Theory (IRT) model is a dramatically important model in educational and psychological measurement. There are two kinds of parameters in the model — item parameters and ability parameters. Nowadays, a commonly used method for estimating item parameters of IRT model is given by Woodruff and Hanson (1997). They treated the ability parameter θ as missing and applied EM Algorithm for finite mixture to estimate item parameters under the condition that the examinees' responses are complete. Here, we extend the Woodruff's method to deal with incomplete response data. That is, we keep the incomplete response cases and regard missing response data as “missing” like θ and then apply EM Algorithm. In our simulation study, we compare the relative performance of the missing data treatment method of us with that of the software BILOG-MG under different sample size and missing ratio. The simulation results show that our new method can obtain better estimation than BILOG-MG in most cases.

Keywords: IRT, missing response, EM algorithm.

AMS Subject Classification: 62N02.