文章编号:1000-5862(2015)06-0623-08

测验 Q 矩阵的修正方法及其比较研究

宋丽红[1],汪文义[2],丁树良[2]
(1. 江西师范大学初等教育学院,江西 南昌 330022;
2. 江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:在认知诊断评估中,构建正确测验 Q 矩阵十分关键,但比较困难. 该文将确定性输入噪音与门模型下 3 种在线标定方法(极大似然估计方法,边际极大似然估计方法和交差方法)用于测验 Q 矩阵修正,并与 δ 方法,γ 方法和最小残差平方和方法进行比较. 采用模拟研究验证和比较各方法的表现,研究结果显示:边际极大似然估计方法表现良好,交差方法次之;项目所考查的属性数目是影响 δ 方法和 γ 方法的表现.

关键词:认知诊断评估;Q 矩阵;确定性输入噪音与门模型;EM 算法;在线标定方法
中图分类号:B 841.7　　**文献标志码**:A　　**DOI**:10.16357/j.cnki.issn1000-5862.2015.06.15

0 引言

认知诊断评估主要用于测量被试的知识结构和加工技能(简称属性). 认知诊断评估的终极目标,是服务于学习和学习进程的评估. 构建 Q 矩阵,是实现认知诊断评估十分关键的步骤之一. Q 矩阵描述了测验项目与属性之间的关联关系,Q 矩阵及其正确性直接决定测验的信度和效度. 本文讨论的 Q 矩阵. 由元素 0 和 1 组成 Q 矩阵也可以是非负整数的多值形式[1-2],它能为细化区分属性掌握程度的差异和简化属性之间关系的描述等.

测验 Q 矩阵是项目与潜在属性之间关系的结构化表征,成为构想效度(construct validity)的主要证据和诊断分类的基础. 因此,如何构建正确的测验 Q 矩阵是绝大多数认知诊断方法要解决的首要问题. 许多研究显示正确构建测验 Q 矩阵比较困难和复杂[3-5],并且错误构建测验 Q 矩阵会带来严重的后果[6-7],比如导致项目参数估计和被试分类不准确等. 所以,Q 矩阵的修正方法成为近年来关注的焦点.

确定性输入噪音与门(deterministic inputs,noisy "and" gate,DINA)模型[8-9]下有多种 Q 矩阵的修正或标定方法:δ 方法[10]、γ 方法[11]、最小残差平方和方法[12]、在线标定方法及其推广方法等[13-22]. 这些方法主要是基于 EM 算法和基于联合极大似然估计(JML)思想,比较容易实现,故可将 3 种在线标定方法用于认知诊断测验 Q 矩阵修正. 由于 DINA 模型下 Q 矩阵修正方法研究相对比较独立,鲜有研究综合比较这些方法的表现,本文将通过模拟研究比较各方法的表现.

1 确定性输入噪音与门模型及其 EM 算法

1.1 确定性输入噪音与门模型

在介绍模型之前,先约定文中使用的基本记号:N 为被试人数,M 为项目个数,K 为属性个数;T 为所有不同的知识状态个数,若属性相互独立,则 $T=2^K$,否则由属性层级结构确定[23];$\boldsymbol{\alpha}_c$ 为第 c 类知识状态,其中 $c=1,2,\cdots,T$,不妨设第 1 类知识状态为属性完全没有掌握的知识状态;$X_i=(X_{i1},X_{i2},\cdots,X_{iM})$ 为被试 i 在 M 个项目上的观察反应模式,其中 $i=1,2,\cdots,N$;$\boldsymbol{q}_j$ 为测验 Q 矩阵中的列向量,表示项目 j 所考查的 K 维属性向量,若项目 j 考察了属性 k,则 $q_{kj}=1$,否则 $q_{kj}=0$,其中 $j=1,2,\cdots,M,k=1,2,\cdots,K$;$\boldsymbol{\eta}_c=(\eta_{c1},\eta_{c2},\cdots,\eta_{cM})$ 表示知识状态为 $\boldsymbol{\alpha}_c$ 的被试在测验 Q 矩阵下的理想反应模式,

收稿日期:2015-05-28
基金项目:国家自然科学基金(31500909,31360237,31300876,31160203,31100756,30860084),教育部人文社会科学研究青年基金(13YJC880060)和江西省教育科学 2013 年度一般课题(13YB032)资助项目.
作者简介:宋丽红(1981-),女,江西新干人,讲师,博士,主要从事教育和心理测量方面研究.

其中 $\eta_{cj}=\prod_{k=1}^{K}\alpha_{kc}^{q_{kj}}$. 因项目属性向量也是知识状态，$\boldsymbol{q}_j$可能为除第 1 类知识状态以外的任一类 $\boldsymbol{\alpha}_c$.

本文主要介绍 DINA 模型[8-9] 下的 Q 矩阵修正方法. DINA 模型是研究较多的非补偿的认知诊断模型，其项目反应函数为

$$P_j(\boldsymbol{\alpha}_i)=P(X_{ij}=1\mid\boldsymbol{\alpha}_i)=g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}=\begin{cases}g_j, & \text{若 } \eta_{ij}=0\Leftrightarrow\boldsymbol{\alpha}_i'\boldsymbol{q}_j<\boldsymbol{q}_j'\boldsymbol{q}_j,\\ 1-s_j, & \text{若 } \eta_{ij}=1\Leftrightarrow\boldsymbol{\alpha}_i'\boldsymbol{q}_j=\boldsymbol{q}_j'\boldsymbol{q}_j,\end{cases}$$

其中 $P_j(\boldsymbol{\alpha}_i)$ 为给定知识状态 $\boldsymbol{\alpha}_i$ 的被试 i 在项目 j 上的正确作答概率，$\eta_{ij}=\prod_{k=1}^{K}\alpha_{ki}^{q_{kj}}$ 为知识状态 $\boldsymbol{\alpha}_i$ 的被试在项目 j 上的理想反应，s_j 和 g_j 分别为项目 j 的失误参数和猜测参数.

1.2 DINA 模型参数估计的 EM 算法

可以采用 EM 或 MCMC 算法进行 DINA 模型参数估计[24-25]，本文仅介绍下面要用到 EM 算法. 在局部独立或条件独立假设下，即同一个被试在各个项目上的作答反应相互独立情况下，知识状态为 $\boldsymbol{\alpha}_i$ 的被试 i 的得分向量 $\boldsymbol{X}_i$ 的似然函数为

$$L(\boldsymbol{X}_i\mid\boldsymbol{\alpha}_i)=\prod_{j=1}^{M}P_j(\boldsymbol{\alpha}_i)^{X_{ij}}(1-P_j(\boldsymbol{\alpha}_i))^{1-X_{ij}}.$$

给定知识状态的先验分布为 $P(\boldsymbol{\alpha}_c)$，$c=1,2,\cdots,T$，由贝叶斯定理可以得到知识状态的后验分布为

$$P(\boldsymbol{\alpha}_c\mid\boldsymbol{X}_i)=L(\boldsymbol{X}_i\mid\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)/\Big(\sum_{l=1}^{T}L(\boldsymbol{X}_i\mid\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)\Big).$$

如通过简单随机取样，在被试相互独立情况下，得分矩阵 $\boldsymbol{X}=(X_{ij})$ 的似然函数为

$$L(\boldsymbol{X}\mid\boldsymbol{\alpha})=\prod_{i=1}^{N}L(\boldsymbol{X}_i\mid\boldsymbol{\alpha}_i)=\prod_{i=1}^{N}\prod_{j=1}^{M}P_j(\boldsymbol{\alpha}_i)^{X_{ij}}\cdot(1-P_j(\boldsymbol{\alpha}_i))^{1-X_{ij}}.$$

将所有被试看成是取自同一总体且给定先验分布 $P(\boldsymbol{\alpha}_c)$，可得边际似然函数为 $L(\boldsymbol{X})=\prod_{i=1}^{N}L(\boldsymbol{X}_i)=\prod_{i=1}^{N}\sum_{c=1}^{T}L(\boldsymbol{X}_i\mid\boldsymbol{\alpha}_c)P(\boldsymbol{\alpha}_c)$.

对边际似然函数求自然对数，代入 DINA 模型的项目反应函数，然后分别对 s_j 和 g_j 参数求一阶偏导，并令 2 个偏导为 0，通过解二元一次方程组，可以得到 s_j 和 g_j 参数的估计值: $\hat{s}_j=(N_j^{(1)}-R_j^{(1)})/N_j^{(1)}$，$\hat{g}_j=R_j^{(0)}/N_j^{(0)}$，其中人工数据 $N_j^{(0)}=\sum_{\alpha_c:\eta_{cj}=0}N_c$ 和 $N_j^{(1)}=\sum_{\alpha_c:\eta_{cj}=1}N_c$ 分别表示项目 j 上理想反应为 0 或 1 的期望人数，$N_c=\sum_{i=1}^{N}P(\boldsymbol{\alpha}_c\mid\boldsymbol{X}_i)$ 表示所有被试中知识状态为 $\boldsymbol{\alpha}_c$ 的期望人数; $R_j^{(0)}=\sum_{\alpha_c:\eta_{cj}=0}R_{jc}$ 和 $R_j^{(1)}=\sum_{\alpha_c:\eta_{cj}=1}R_{jc}$ 分别表示项目 j 上理想反应为 0 或 1 且正确作答项目 j 的期望人数，其中 $R_{jc}=\sum_{i=1}^{N}P(\boldsymbol{\alpha}_c\mid\boldsymbol{X}_i)^{X_{ij}}$ 表示所有被试中知识状态为 $\boldsymbol{\alpha}_c$ 且正确作答项目 j 的期望人数.

在估计出项目参数之后，可使用极大似然估计(MLE)、最大后验估计(MAP)和期望后验估计(EAP)方法估计被试的知识状态，估计公式分别为

$$\hat{\boldsymbol{\alpha}}_i=\arg\max_{\boldsymbol{\alpha}_c}\{L(\boldsymbol{X}_i\mid\boldsymbol{\alpha}_c)\},$$

$$\hat{\boldsymbol{\alpha}}_i=\arg\max_{\boldsymbol{\alpha}_c}\{P(\boldsymbol{\alpha}_c\mid\boldsymbol{X}_i)\},$$

$$\hat{\boldsymbol{\alpha}}_{ik}=\begin{cases}1, & \text{if } \alpha'_{ik}\geqslant 0.5,\\ 0, & \text{otherwise},\end{cases}$$

其中 $\alpha'_{ik}=\sum_{c=1}^{T}p(\boldsymbol{\alpha}_c\mid\boldsymbol{X}_i)\alpha_{kc}$，$\alpha'_{ik}$ 取值 [0,1] 之间某个值. 为得到 0-1 二值数值，通常根据属性划界分数决定属性掌握与未掌握，如上面 EAP 估计中取划界分数为 0.5. 也有将 α_{ik}' 分为 3 类 [0,0.4，0.6] 代表属性不确定区间; [0,0.4) 代表属性未掌握区间; (0.6,1] 代表属性掌握区间.

2 Q 矩阵修正方法

本文主要关注基于 EM 算法和基于联合极大似然估计思想的 Q 矩阵修正方法. 给定初始 Q 矩阵和得分矩阵，这类方法的基本步骤可以归纳如下:

(i) 给定初始或更新的 Q 矩阵，使用一次 EM 算法估计项目参数 $\hat{s}_j$ 和 $\hat{g}_j$，然后估计被试知识状态 $\hat{\boldsymbol{\alpha}}_i$ 及其后验分布 $p(\boldsymbol{\alpha}_c\mid\boldsymbol{X}_i)$ 和属性掌握概率 α_{ik}'，或者不估计项目参数而使用其他方法直接对被试知识状态进行分类;

(ii) 选择每个项目 j，采用 Q 矩阵修正方法更新 Q 矩阵中第 j 列属性向量 $\boldsymbol{q}_j$;

(iii) 重复以上 2 个步骤，直到收敛准则满足为止.

在第(ii)步中 $\boldsymbol{q}_j$ 的估计可以使用下面详细介绍的 δ 方法[10]，γ 方法[11]，最小残差平方和方法[12] 和 3 种在线标定方法(极大似然估计方法，边际极大似然估计方法和交差方法)的推广方法. 本文使用相对收敛准则的停止准则: $|\log(L(\boldsymbol{X}))^{(s+1)}-\log(L(\boldsymbol{X}))^{(s)}|/|\log(L(\boldsymbol{X}))^{(s)}|<0.001$.

2.1　δ 方法

de la Torre[10] 提出了基于 DINA 模型的 δ 方法修正 Q 矩阵,并对分数减法数据[26] 的 Q 矩阵进行了分析. δ 方法的基本思想是: 选择属性向量 $\boldsymbol{q}_j = \boldsymbol{\alpha}_c (c \neq 1)$ 使掌握项目 j 所考查的所有属性的被试组与没有完全掌握项目 j 所考查属性的被试组在项目 j 上正确作答概率之差最大化. δ 方法的数学表达式为: $\boldsymbol{q}_j = \arg\max_{\boldsymbol{a}_c} [P(X_j = 1 \mid \eta_{c'c} = 1) - P(X_j = 1 \mid \eta_{c'c} = 0)] = \arg\max_{\boldsymbol{a}_c}(\delta_{jc})$,其中 $c = 2,3,\cdots,T$, $c' = 1,2,\cdots,T$, $\eta_{c'c} = \prod_{k=1}^{K} \alpha_{kc}^{\alpha_{kc'}}$. 在 DINA 模型下,因 $P(X_j = 1 \mid \eta_{c'c} = 1) = 1 - s_j$ 和 $P(X_j = 1 \mid \eta_{c'c} = 0) = g_j$,使 δ_{jc} 最大化就是最大化 $1 - s_j - g_j$,即最小化失误参数 s_j 和猜测参数 g_j 之和. 必须注意的是失误参数 s_j 和猜测参数 g_j 是估计第 j 个项目的属性向量 $\boldsymbol{q}_j$ 的充分条件,但是这并不是模型—资料拟合的必要条件.

对于项目 j 而言,使用 δ_{jc} 指标估计属性向量 $\boldsymbol{q}_j$,一种最为直接的方式是使用穷举法(exhaustive search algorithm),试探每一种可能的属性向量(即除去全零向量的知识状态),搜索空间的元素个数为 $T - 1$. de la Torre 认为搜索空间的元素个数与属性数成指数式增长,穷举法只适合属性数比较小的情况. 因此,de la Torre 提出了一种序贯搜索算法(sequential search algorithm),基本步骤如下:

(i) 试探考查一个属性的所有属性向量,找出使 δ_{jc} 最大的属性 k_1,并记此时达到的最大的 δ_{jc} 值为 $\delta_j^{(1)}$,同时更新 $\boldsymbol{q}_j$ 的第 k_1 个分量值为 1,其余分量为 0,即认为项目 j 考查了属性 k_1,进入(ii).

(ii) 试探考查 2 个属性的所有属性向量(属性 k_1 和其他任一属性),找出使 δ_{jc} 最大的属性 k_2,并记此时最大的 δ_{jc} 值为 $\delta_j^{(2)}$. 若 $\delta_j^{(2)} - \delta_j^{(1)} > \varepsilon$,更新 $\boldsymbol{q}_j$ 的第 k_2 个分量值为 1,即认为项目 j 考查了属性 k_1 和 k_2,进入(iii);否则不更新 $\boldsymbol{q}_j$,结束搜索.

(iii) 依次试探考查 $s(s = 3,4,\cdots,K)$ 个属性的所有属性向量($\boldsymbol{q}_j$ 中已考查的 $s - 1$ 个属性和其他任一属性),找出使 δ_{jc} 最大的属性 k_s,并记此时最大的 δ_{jc} 值为 $\delta_j^{(s)}$. 若 $\delta_j^{(s)} - \delta_j^{(s-1)} > \varepsilon$,更新 $\boldsymbol{q}_j$ 的第 k_s 个分量值为 1,即认为项目 j 考查了属性 $k_1,k_2,\cdots,k_s$,继续步骤(iii);否则不更新 $\boldsymbol{q}_j$,结束搜索.

当 K 比较大时,序贯搜索算法的搜索次数为 $(K^2 + K)/2$,远小于属性间相互独立条件下穷举法的搜索次数 $T - 1 = 2^K - 1$. 如果项目只考查了 K_j 个属性,理论上试探完考查 $K_j + 1$ 个属性的所有属性向量时,算法就会达到终止条件,因此总的搜索次数仅为 $K + (K - 1) + \cdots + (K - K_j) = (K_j + 1)K - (K_j^2 + K_j)/2$.

de la Torre 使用 EM 算法的估计结果计算 δ_{jc},这里的 c 指示序贯搜索算法搜索的属性向量为 $\boldsymbol{\alpha}_c$. 当试探项目 j 的属性向量 $\boldsymbol{a}_c$ 时,$\delta_{jc} = 1 - \hat{s}_{j_c} - \hat{g}_{j_c}$,其中 $\hat{s}_{j_c} = (N_{j_c}^{(1)} - R_{j_c}^{(1)})/N_{j_c}^{(1)}$,$\hat{g}_{j_c} = R_{j_c}^{(0)}/N_{j_c}^{(0)}$,且 $N_{j_c}^{(0)} = \sum_{\alpha_c:\eta_{c'c}=0} N_{c'}$,$N_{j_c}^{(1)} = \sum_{\alpha_c:\eta_{c'c}=1} N_{c'}$,$N_{c'} = \sum_{i=1}^{N} P(\boldsymbol{\alpha}_{c'} \mid \boldsymbol{X}_i)$,$R_{j_c}^{(0)} = \sum_{\alpha_c:\eta_{c'c}=0} R_{jc'}$,$R_{j_c}^{(1)} = \sum_{\alpha_c:\eta_{c'c}=1} R_{jc'}$ 和 $R_{jc'} = \sum_{i=1}^{N} P(\boldsymbol{\alpha}_{c'} \mid X_i)^{X_{ij}}$. 在 EM 算法中是用 $\boldsymbol{q}_j$ 计算人工数据,而序贯搜索算法试探项目 j 的属性向量 $\boldsymbol{\alpha}_c$ 时,只需令 $\boldsymbol{q}_j = \boldsymbol{\alpha}_c$ 计算人工数据并估计项目参数得到 δ_{jc}. 由于测验 Q 矩阵误指会影响项目参数估计和 $P(\boldsymbol{\alpha}_{c'} \mid \boldsymbol{X}_i)$ 的估计,从而影响 δ_{jc},因此需要设置阈值 ε 以结束搜索. 设置的阈值 ε 越小,添加到 $\boldsymbol{q}_j$ 中的属性会越多;否则会越少.

2.2　γ 方法

涂冬波等[11] 提出 γ 方法用于 Q 矩阵修正. 该方法的基本假设是: 若项目 j 考查了属性 k,则掌握属性 k 的被试组在项目 j 的得分应高于未掌握属性 k 的被试组的得分;若项目 j 未考查属性 k,则掌握属性 k 的被试组与未掌握属性 k 的被试组在项目 j 的得分相当. γ 方法的主要步骤如下:

(i) 采用 DINA 模型的 EM 算法估计项目参数($\hat{s}_j$ 和 $\hat{g}_j$)及被试对每个属性的掌握概率 α_{ik}'(如采用 EAP 方法估计);

(ii) 对 Q 矩阵中满足一定条件的元素(其余元素不变)逐个进行修正,修正公式为

$$q_{jk} = \begin{cases} 0, \text{if } \hat{g}_j > \text{临界值且 } ES_{jk} < 0.2, \\ 1, \text{if } \hat{s}_j > \text{临界值且 } ES_{jk} \geqslant 0.2, \end{cases}$$

其中效应值

$$ES_{jk} = \sum_{i=1}^{N} I(\alpha_{ik}' > 0.6)^{X_{ij}} / \left(\sum_{i=1}^{N} I(\alpha_{ik}' > 0.6)\right) - \sum_{i=1}^{N} I(\alpha_{ik}' < 0.4)^{X_{ij}} / \left(\sum_{i=1}^{N} I(\alpha'_{ik} < 0.4)\right),$$

$j = 1,2,\cdots,M$, $k = 1,2,\cdots,K$.

γ 方法并不迭代估计被试属性掌握概率和修正 Q 矩阵元素. γ 方法对临界值敏感,根据涂冬波等[11] 的研究结果,临界值取 0.2 较为适宜,本文实验中临界值取 0.2.

2.3 最小残差平方和方法(RSS)

Chiu Chiayi[12]提出了基于最小残差平方和(residual sum of squares,RSS)方法进行 Q 矩阵修正.最小残差平方和方法是一种用于估计属性向量的非参数方法,它与知识状态的非参数分类方法[27]相结合,实现属性向量与知识状态迭代估计.给定初始 Q 矩阵,记为 $Q^{(0)}$,最小残差平方和方法的主要步骤如下:

(i)初始化待修正的项目集 $S^{(0)}=\{1,2,\cdots,M\}$;

(ii)根据 $Q^{(0)}$ 计算各知识状态 $\hat{\boldsymbol{\alpha}}_{c'}$ 的理想反应 η_{cj},采用非参数分类方法之海明距离(或加权海明距离)[27]方法估计被试的潜在知识状态:$\hat{\boldsymbol{\alpha}}_i=\arg\min_{\boldsymbol{\alpha}_{c'}}\left\{\sum_{i=1}^{M}|X_{ij}-\eta_{cj}|\right\}$;

(iii)基于 $Q^{(0)}$ 和 $\hat{\boldsymbol{\alpha}}_i$,计算各个被试在各个项目上的理想反应 η_{ij},再计算 $S^{(0)}$ 中各个项目的残差平方和:$RSS_j=\sum_{i=1}^{N}(X_{ij}-\eta_{ij})^2$.选择残差平方和最大的项目 j 进入(iv)以进行属性向量修正;

(iv)更新 $Q^{(0)}$ 中项目 j 的属性向量为:$\boldsymbol{q}_j=\arg\min_{\boldsymbol{\alpha}_c;\,c\neq 1}\sum_{i=1}^{N}(X_{ij}-\eta_{ic})^2$ 且令 $S^{(0)}=S^{(0)}-\{j\}$(集合差运算),重复(ii)和(iv),直至 $S^{(0)}$ 为空集,进入步骤(v);

(v)重复步(i)~(iv),直至所有项目的残差平方和不变为止.

2.4 极大似然估计方法(MLE)

极大似然估计方法可看成是对属性向量的一种条件估计方法.基于初始 Q 矩阵和得分矩阵,使用一次 EM 算法估计测验项目参数 $\hat{s}_j$ 和 $\hat{g}_j$,并由测验项目参数估计出被试知识状态 $\hat{\boldsymbol{\alpha}}_i$.由 DINA 模型的项目反应函数,给定被试知识状态 $\hat{\boldsymbol{\alpha}}_i$ 与项目参数 $\hat{s}_j$ 和 $\hat{g}_j$,可计算项目 j 的属性向量为 $\boldsymbol{\alpha}_c$ 条件下的项目反应概率为 $P_{j_c}(\hat{\boldsymbol{\alpha}}_i)$.在被试相互独立情况下,令项目 j 上得分向量 $\boldsymbol{X}_j$ 的似然函数为 $L(\boldsymbol{X}_j|\boldsymbol{\alpha}_c)$,由极大似然估计方法估计项目 j 的 $\boldsymbol{q}_j$ 为 $\boldsymbol{q}_j=\arg\max_{\boldsymbol{\alpha}_c;\,c\neq 1}L(\boldsymbol{X}_j|\alpha_c)=\arg\max_{\boldsymbol{\alpha}_c;\,c\neq 1}\prod_{i=1}^{N}P_{j_c}(\hat{\boldsymbol{\alpha}}_i)^{X_{ij}}(1-P_{j_c}(\hat{\boldsymbol{\alpha}}_i))^{1-X_{ij}}$.

为计算方便,可统计所有被试中每种知识状态 $\hat{\boldsymbol{\alpha}}_{c'}$ 的被试正确与错误作答项目 j 的频数 r_{cj} 和 w_{cj},并将上式转化成对数似然函数为

$$\boldsymbol{q}_j=\arg\max_{\boldsymbol{\alpha}_c;\,c\neq 1}\sum_{c'=1}^{T}\left(r_{cj}\log P_{j_c}(\hat{\boldsymbol{\alpha}}_{c'})+w_{cj}\log(1-P_{j_c}(\hat{\boldsymbol{\alpha}}_{c'}))\right).$$

2.5 边际极大似然估计方法(MMLE)

边际极大似然估计方法仅设定被试的知识状态为随机向量并指定其先验分布,被广泛应用于部分贝叶斯模型(partially Bayesian models)参数估计[28].观察数据的似然函数涉及的未知参数包括项目参数、被试知识状态和项目属性向量.在部分贝叶斯模型框架下,视被试的知识状态为随机向量(以后验分布为先验分布)而固定当前估计的项目参数(并没有考虑项目参数的估计误差或后验分布),对项目属性向量进行条件估计.因此,边际极大似然估计方法估计项目 j 的 $\boldsymbol{q}_j$ 为

$$\boldsymbol{q}_j=\arg\max_{\boldsymbol{\alpha}_c;\,c\neq 1}\prod_{i=1}^{N}\sum_{c'=1}^{T}P_{j_c}(\boldsymbol{\alpha}_{c'})^{X_{ij}}(1-P_{j_c}(\boldsymbol{\alpha}_{c'}))^{1-X_{ij}}P(\boldsymbol{\alpha}_{c'}|\boldsymbol{X}_i).$$

2.6 交差方法(I&D)

考虑项目内属性之间没有补偿作用,先考虑理想反应情况.对项目 j 正确反应的被试的知识状态所对应的属性集合(以下简称为知识状态并记为 $\boldsymbol{\alpha}_{c'}$),必有项目 j 所包含的属性集合 $\boldsymbol{q}_j$ 是 $\boldsymbol{\alpha}_{c'}$ 所包含的属性集合的子集,记为 $\boldsymbol{q}_j\subseteq\boldsymbol{\alpha}_{c'}$.记所有答对项目 j 的知识状态的属性集合 $\boldsymbol{\alpha}_{c'}$ 交集为:$upper=\bigcap\{\boldsymbol{\alpha}_c:\boldsymbol{q}_j=\boldsymbol{\alpha}_c,c\neq 1,\boldsymbol{q}_j\subseteq\boldsymbol{\alpha}_{c'}\}$,由理想反应模式的定义及集合论可知,$\boldsymbol{q}_j\subseteq upper$.而对项目 j 错误反应的所有被试的知识状态 $\boldsymbol{\alpha}_{c'}$ 都不可能包含 $\boldsymbol{q}_j$,错误反应的被试的知识状态集中有一部分知识状态 $\boldsymbol{\beta}$ 是 $\boldsymbol{q}_j$ 的真子集,将满足这一条件的 $\boldsymbol{\beta}$ 作并运算,并将运算结果记为 $lower$,由于并运算可能会使得集合所包含的元素增加,于是 $lower$ 可能更接近 $\boldsymbol{q}_j$.特别是对项目 j 反应的人数较多时,有 $lower\subseteq\boldsymbol{q}_j\subseteq upper$.

但是以上是对理想反应情况的讨论,实际反应中存在猜测与失误.故设定一个指标,若具有某种知识状态 $\boldsymbol{\alpha}_{c'}$ 的所有被试(人数为 $n_{cj}=r_{cj}+w_{cj}$)对某题(可以抽象为 Q 矩阵中某一列 $\boldsymbol{q}_j$)的答对比率($p_{cj}=r_{cj}/n_{cj}$)高于答错比率($1-p_{cj}$),即 $p_{cj}>1-p_{cj}$,则认为知识状态为 $\boldsymbol{\alpha}_{c'}$ 的被试掌握了正确回答项目 j 所需的所有属性 $\boldsymbol{q}_j$,被试应该答对该项目,答错只是因为失误.从集合论的观点来看,$\boldsymbol{q}_j$ 所包含的属性集合是 $\boldsymbol{\alpha}_{c'}$ 所包含的属性集合的子集($\boldsymbol{q}_j\subseteq\boldsymbol{\alpha}_{c'}$).如果从偏序关系的观点来看,可以认为 $\boldsymbol{q}_j$ 是 $\boldsymbol{\alpha}_{c'}$ 的下界,即令 $L_{\boldsymbol{\alpha}_{c'}}=\{\boldsymbol{q}_j|\forall\,\boldsymbol{q}_j=\boldsymbol{\alpha}_c;c\neq 1$ 且 $\boldsymbol{q}_j\subseteq\boldsymbol{\alpha}_c\}$,其中 $\boldsymbol{q}_j\subseteq\boldsymbol{\alpha}_{c'}$ 表示向量 $\boldsymbol{q}_j$ 中元素均小于或等于 $\boldsymbol{\alpha}_{c'}$ 中对应元素,知 $L_{\boldsymbol{\alpha}_{c'}}$ 为 $\boldsymbol{\alpha}_{c'}$ 的下界,则 $\boldsymbol{q}_j\in L_{\boldsymbol{\alpha}_{c'}}$.若 $p_{cj}\leqslant 1-p_{cj}$,$\boldsymbol{q}_j$ 所包含的属性集合不是 $\boldsymbol{\alpha}_{c'}$ 所包含

的属性集合的子集,即 $q_j \notin L_{\alpha_{c'}}$,被试本不应该答对,答对只是因为猜测.

由于存在失误和猜测和知识状态估计存在误差,某些知识状态答对与答错的比率可能比较接近,如 0.51 或 0.49 时. 对于这种情况,交差方法将优先选择答对与答错的比率或答错与答对的比率大的作交差运算,当得到的属性向量的候选空间仅剩一个时,就不再考虑其他知识状态. 交差方法采用下述步骤来实现: 设 q_j 的候选集为 $C_j = \{\boldsymbol{\alpha}_c : c \neq 1\}$;对每个知识状态 $\boldsymbol{\alpha}_{c'}$ 计算 $\max(p_{cj}/(1-p_{cj}), (1-p_{cj})/p_{cj})$ (分母为 0 时,r_{cj} 或 w_{cj} 设为 1),按从大到小排序; 按从大到小顺序依次取出 $\boldsymbol{\alpha}_{c'}$ 若 $p_{cj} > 1 - p_{cj}$,$C_j \Leftarrow C_j \cap L_{\boldsymbol{\alpha}_{c'}}$,否则 $C_j \Leftarrow C_j - L_{\boldsymbol{\alpha}_{c'}}$,直到 C_j 只有一个元素,即 $\{q_j\} \Leftarrow C_j$ (对所有知识状态均已经循环后仍包含有多个元素时判错). 由于估计方法使用了集合的交运算和差运算,故称之为交差方法.

3　模拟研究

模拟研究的主要目的是在 DINA 模型下验证和比较用于测验 Q 矩阵修正的 6 种方法的表现,重点考虑影响修正方法表现的主要因素.

3.1　研究设计

考察 5 个属性的测验. 考虑可能影响 Q 矩阵修正的 4 个主要因素: 知识状态分布(2 个水平)、样本量(3 个水平)、测验项目质量(2 个水平)和测验 Q 矩阵元素误指率(5 个水平). 实验条件组合数共 $2 \times 3 \times 2 \times 5 = 60$ 种,每种实验条件重复 200 次. 知识状态分布服从多元离散均匀分布和多元正态阈值模型(multivariate normal threshold model)[12]. 多元正态阈值模型先通过模拟多元正态分布的随机数 $\theta \sim MVN(0, \Sigma)$,其中协方差阵 Σ 中主对角线元素全为 1,其他元素全设为 0.5. 然后根据属性阈值对连续的随机数 θ_{ik} 进行 0-1 化,公式如下:

$$\alpha_{ik} = \begin{cases} 1, \text{if } \theta_{ik} \geqslant \Phi^{-1}(k/(K+1)), \\ 0, \text{otherwise}, \end{cases}$$

其中 $i = 1, 2, \cdots, N, k = 1, 2, \cdots, K, \Phi^{-1}$ 为标准正态分布的累积分布函数的逆函数. 根据相应的知识状态分布,分别模拟 300,500 和 1 000 个被试. DINA 模型下测验项目参数服从均匀分布 $U(0.05, 0.25)$ 或 $U(0.05, 0.40)$. 根据 Q 矩阵理论[29-30],约简 Q 矩阵作为真实测验 Q 矩阵. 使用真实测验 Q 矩阵模拟得分矩阵. 为考虑 Q 矩阵误指,基于真实测验 Q 矩阵,随机模拟生成 4 种误指水平的测验 Q 矩阵,元素误指率分别为 0.1,0.2,0.3,0.4. 加上正确测验 Q 矩阵,误指率为 0,共 5 个水平的初始 Q 矩阵.

3.2　分析步骤

对于同一批得分矩阵和初始 Q 矩阵,分别使用 δ 方法(阈值 ε 设置为 0.2)、γ 方法(临界值设置为 0.2)、最小残差平方和方法、极大似然估计方法、边际极大似然估计方法和交差方法对初始 Q 矩阵进行验证或修正,得到修正的测验 Q 矩阵. 再计算修正的测验 Q 矩阵所有元素的返真率. 对于相同条件下重复 200 次的返真率,基于成对数据的 t 检验,设置显著性水平为 0.05. 原假设 H_0: 平均返真率最高的方法与任一其他方法所得返真率相等. 另外还进行了威尔科克森符号秩检验(Wilcoxon signed rank test),类似结果未列出.

为了考察各方法对 Q 矩阵中项目属性向量的误指类型的影响,分别统计了初始 Q 矩阵中 4 种类型项目的属性向量元素的返真率. 这 4 种类型分别是: 多指(over-specified)元素(属性向量中有些 0 元素误指为 1,0→1)、少指(under-specified)元素(属性向量中有些 1 元素误指为 0,1→0)、混合误指(1→0&0→1)和没有误指(none). 为了考虑项目所考查的属性数的影响,还统计了考查不同属性数项目的属性向量元素的返真率.

3.3　实验结果

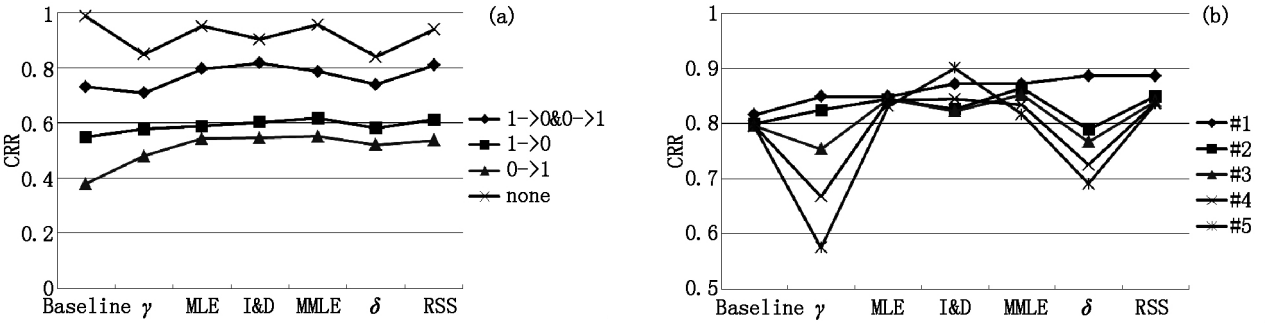
表 1 给出了各实验条件下各方法修正的测验 Q 矩阵元素的返真率均值,其中粗体表示最高返真率,粗体加斜体表示在显著性水平 0.05 下与最高返真率无显著差异的结果. 图 2 汇总了各方法对 4 种类型或考查不同属性数项目的属性向量元素的返真率均值. 返真率的标准差来列出,其范围在 0 和0.12之间. 下面分析各因素对返真率的影响: (i) 知识状态分布的影响. 表 1 显示在 2 种分布下各方法均可提高初始 Q 矩阵的质量. 由于 EM 算法中知识状态的先验分布匹配多元离散均匀分布,多元离散均匀分布结果优于多元正态阈值模型. 另一个原因,多元正态阈值模型属性间中等相关且属性掌握的难度递增,导致了许多知识状态对应的人数太少,从而难于准确标定一些项目的属性向量; (ii) 样本量的影响. 表 1 显示样本量从 500 增加到 1 000 时,Q 矩阵元素的返真率明显提高. 在样本量为 1 000 时,许多条件下的 Q 矩阵元素的返真率高于 0.9 甚至接近 1. 在误指率较低的条件下,DINA 模型下样本量1 000 可以获得较高的返真率. 但在误指率较高时,增加样本量也较难提高返真率; (iii) 项目质量的影响. 当猜测和失误参数较小时,知识状态的判准率较高,表 1 显示此时 Q 矩阵元素的返真率较高; (iv) 误指率的

影响. 误指率较低和中等时,Q 矩阵元素的返真率较高,样本量 300 时除外. 而当误指率达到 0.4 时,Q 矩阵元素的返真率不佳; (v) 误指类型的影响. 图 2 (a) 汇总了所有实验条件下各方法在各类型项目上的返真率均值. 与基准值(初始未修正 Q 矩阵) 比较发现,各方法基本上可以提高各类型项目上的返真率,只是对于少指属性项目的返真率的提高幅度较小; (vi) 属性数的影响. 图 2(b) 汇总了所有实验条件下各方法在不同属性数项目的返真率均值. 极大似然估计方法、边际极大似然估计方法、最小残差平方和方法、交差方法的表现,受属性数的影响较小. 而 γ 方法和 δ 方法在所含属性数较多项目上表现不佳; (vii) 各方法的表现. 表 1 显示,边际极大似然估计方法、交差方法、极大似然估计方法、最小残差平方和方法优于 γ 方法和 δ 方法. 总的来说,边际极大似然估计方法表现良好; 交差方法表现次之,特别在误指率较高或均匀分布下表现较好.

表 1 各实验条件下各方法修正的测验 Q 矩阵元素的返真率均值

		知识状态分布											
		多元离散均匀分布						多元正态阈值模型					
		修正方法						修正方法					
N	ER	γ	MLE	I&D	MMLE	δ	RSS	γ	MLE	I&D	MMLE	δ	RSS
$s_j \sim U(0.05,0.25)$, $g_j \sim U(0.05,0.25)$													
300	0.00	0.935	0.997	0.994	0.998	0.987	0.999	0.907	0.959	0.902	0.976	0.795	0.976
	0.10	0.896	0.986	0.993	0.989	0.973	0.996	0.844	0.909	0.859	0.921	0.776	0.923
	0.20	0.865	0.962	0.986	0.980	0.955	0.988	0.785	0.851	0.812	0.858	0.743	0.844
	0.30	0.808	0.878	0.933	0.924	0.882	0.898	0.707	0.762	0.737	0.769	0.683	0.753
	0.40	0.664	0.679	0.718	0.705	0.688	0.696	0.616	0.645	0.654	0.653	0.613	0.644
500	0.00	0.937	1.00	0.999	1.00	0.995	1.00	0.917	0.977	0.933	0.987	0.805	0.984
	0.10	0.896	0.994	0.999	0.995	0.987	0.999	0.854	0.923	0.891	0.933	0.787	0.937
	0.20	0.863	0.982	0.995	0.990	0.978	0.997	0.794	0.865	0.845	0.876	0.751	0.860
	0.30	0.810	0.899	0.943	0.941	0.897	0.924	0.720	0.784	0.769	0.789	0.701	0.766
	0.40	0.661	0.685	0.745	0.718	0.711	0.714	0.623	0.647	0.661	0.650	0.621	0.650
1 000	0.00	0.936	1.00	1.00	1.00	0.999	1.00	0.915	0.989	0.957	0.994	0.806	0.990
	0.10	0.895	0.998	1.00	0.999	0.993	1.00	0.857	0.937	0.923	0.941	0.788	0.947
	0.20	0.864	0.991	0.998	0.996	0.990	0.996	0.797	0.875	0.873	0.883	0.752	0.879
	0.30	0.818	0.919	0.968	0.974	0.948	0.943	0.722	0.783	0.790	0.786	0.696	0.769
	0.40	0.674	0.699	0.754	0.750	0.739	0.730	0.619	0.648	0.672	0.650	0.617	0.648
$s_j \sim U(0.05,0.40)$, $g_j \sim U(0.05,0.40)$													
300	0.00	0.843	0.982	0.948	0.987	0.931	0.986	0.830	0.949	0.853	0.966	0.800	0.947
	0.10	0.810	0.937	0.925	0.946	0.885	0.943	0.789	0.882	0.807	0.896	0.772	0.883
	0.20	0.764	0.865	0.873	0.885	0.824	0.877	0.736	0.807	0.749	0.818	0.723	0.806
	0.30	0.701	0.759	0.777	0.785	0.736	0.770	0.674	0.721	0.687	0.733	0.672	0.718
	0.40	0.597	0.625	0.650	0.633	0.625	0.644	0.592	0.619	0.622	0.626	0.604	0.626
500	0.00	0.842	0.995	0.979	0.996	0.947	0.995	0.842	0.973	0.907	0.983	0.821	0.966
	0.10	0.807	0.950	0.956	0.955	0.910	0.956	0.801	0.901	0.854	0.910	0.790	0.905
	0.20	0.767	0.882	0.906	0.899	0.851	0.885	0.749	0.826	0.794	0.838	0.742	0.825
	0.30	0.698	0.770	0.802	0.796	0.753	0.782	0.688	0.737	0.724	0.742	0.684	0.736
	0.40	0.601	0.631	0.666	0.640	0.637	0.651	0.605	0.622	0.634	0.629	0.611	0.635
1 000	0.00	0.846	0.999	0.994	0.999	0.963	0.998	0.844	0.986	0.949	0.992	0.828	0.979
	0.10	0.810	0.963	0.975	0.963	0.927	0.966	0.803	0.915	0.893	0.921	0.800	0.914
	0.20	0.770	0.889	0.936	0.912	0.877	0.905	0.755	0.838	0.825	0.848	0.748	0.842
	0.30	0.707	0.773	0.838	0.803	0.782	0.793	0.690	0.742	0.739	0.747	0.684	0.741
	0.40	0.610	0.633	0.678	0.646	0.652	0.655	0.595	0.621	0.637	0.628	0.606	0.642

注: N = 样本量; ER = 误指率; γ = γ 方法; MLE = 极大似然估计方法; I&D = 交差方法; MMLE = 边际极大似然估计方法; δ = δ 方法; RSS = 最小残差平方和方法. 图 2 中缩写含义与之相同.



注: Baseline 表示未修正的初始 Q 矩阵，#1，#2，…，#5 表示项目所考查的属性数.

图 1　各方法对 4 种类型(a) 或不同属性数(b) 项目的属性向量元素的返真率

4　小结与讨论

根据对各修正方法的分析和模拟研究结果发现: 边际极大似然估计方法、交差方法、最小残差平方和方法、极大似然估计方法表现优于 δ 方法和 γ 方法，其中边际极大似然估计方法表现最优; 交差方法表现次之，特别在误指率较高或多元离散均匀分布条件下表现较好. δ 方法和 γ 方法需要设定或选择合适的阈值或临界值，而阈值或临界值可能依赖于项目所含属性数等其他因素.

若仿照回归分析中逐步筛选变量的过程，对 Q 矩阵元素进行修正，效果如何值得进一步考虑. 还有诸多未进行比较的修正方法: 贝叶斯估计方法[3]，该方法结合 Q 矩阵元素的先验知识使用 MCMC 算法对测验 Q 矩阵进行修正，该方法需要预先指定不确定元素; 数据驱动方法[31] 及其改进方法[32]，涉及复杂的 T 矩阵和迭代算法，计算量相当大; 最近喻晓锋等[22] 提出的基于似然比统计量的项目属性向量在线估计方法，该方法对每个新题要调用 EM 算法 $T-1$ 次，计算量仍比较大.

5　参考文献

[1] Chen Jinsong，de la Torre J. A general cognitive diagnosis model for expert-defined polytomous attributes [J]. Applied Psychological Measurement，2013，37(6) : 419-437.

[2] Sun Jia´nan，Xin Tao，Zhang Shumei，et al. A polytomous extension of the generalized distance discriminating method [J]. Applied Psychological Measurement，2013，37 (7) : 503-521.

[3] De Carlo L T. On the analysis of fraction subtraction data: the DINA model，classification，latent class sizes，and the Q-matrix [J]. Applied Psychological Measurement，2011，35(1) : 8-26.

[4] Jang E E. Cognitive diagnostic assessment of L_2 reading comprehension ability: validity arguments for fusion model application to languedge assessment [J]. Language Testing，2009，26(1) : 31-73.

[5] McGlohen M K，Chang Huahua. Combining computer adaptive testing technology with cognitively diagnostic assessment [J]. Behavior Research Methods，2008，40(3) : 808-821.

[6] Im S，Corter J E. Statistical consequences of attribute misspecification in the rule space method [J]. Educational and Psychological Measurement，2011，71(4) : 712-731.

[7] Rupp A A，Templin J. The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model [J]. Educational and Psychological Measurement，2008，68(1) : 78-96.

[8] Junker B，Sijtsma K. Cognitive assessment models with few assumptions，and connections with nonparametric item response theory [J]. Applied Psychological Measurement，2001，25: 258-272.

[9] Haertel E H. Using restricted latent class models to map the skill structure of achievement items [J]. Journal of Educational Measurement，1989，26(4) : 301-321.

[10] dela Torre J. An empirically based method of Q-matrix validation for the DINA model: development and applications [J]. Journal of Educational Measurement，2008，45: 343-362.

[11] 涂冬波，蔡艳，戴海琦. 基于 DINA 模型的 Q 矩阵修正方法 [J]. 心理学报，2012，44(4) : 558-568.

[12] Chiu Chiayi. Statistical refinement of the Q-Matrix in cognitive diagnosis [J]. Applied Psychological Measurement，2013，37(8) : 598-618.

[13] Chen Ping，Xin Tao，Wang Chun，et al. On-line calibration methods for the DINA model with independent attributes in CA-CAT [J]. Psychometrika，2012，77(2) : 201-222.

[14] 陈平，辛涛. 认知诊断计算机化自适应测验中在线标定方法的开发 [J]. 心理学报，2011，43(6) : 710-724.

[15] 陈平，辛涛. 认知诊断计算机化自适应测验中的项目增补 [J]. 心理学报，2011，43(7) : 836-850.

[16] 陈平,张佳慧,辛涛. 在线标定技术在计算机化自适应测验中的应用 [J]. 心理科学进展,2013,21(10):1883-1892.

[17] 汪文义,丁树良. 题库结构对原始题在线属性标定准确性之影响研究 [J]. 心理科学,2012,35(2):452-456.

[18] 汪文义,丁树良,游晓锋. 计算机化自适应诊断测验中原始题的属性标定 [J]. 心理学报,2011,43(8):964-976.

[19] Wang Wenyi, Ding Shuliang, Song Lihong. New *Q*-matrix validation methods and their sensitivity under the DINA model [C]. San Francisco: CA, 2013.

[20] 丁树良,罗芬,汪文义. 认知诊断分类中心的确定 [J]. 心理学探新,2013,33(5):396-401.

[21] 汪文义,宋丽红,丁树良. 基于探索性因素分析的 *Q* 矩阵标定方法 [J]. 江西师范大学学报: 自然科学版,2015,39(2):138-144,170.

[22] 喻晓锋,罗照盛,高椿雷,等. 使用似然比 D^2 统计量的题目属性定义方法 [J]. 心理学报,2015,47(3):417-426.

[23] Leighton J P, Gierl M J, Hunka S M. The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach [J]. Journal of Educational Measurement, 2004, 41(3): 205-237.

[24] de la Torre J. DINA model and parameter estimation: a didactic [J]. Journal of Educational and Behavioral Statistics, 2009, 34(1): 115-130.

[25] de la Torre J, Douglas J. Higher-order latent trait models for cognitive diagnosis [J]. Psychometrika, 2004, 69(3): 333-353.

[26] Tatsuoka K K. Toward an integration of item-response theory and cognitive error diagnosis [C] // Frederiksen N, Glaser R L, Lesgold A M, et al. Diagnostic monitoring of skill and knowledge acquisition [A]. Erlbaum: Hillsdale, 1990: 453-488.

[27] Chiu Chiayi, Douglas J A. A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns [J]. Journal of Classification, 2013, 30: 225-250.

[28] DiBello L V, Roussos L A, Stout W. Review of cognitively diagnostic assessment and a summary of psychometric models [C] // Rao C R, Sinharay S. Handbook of statistics [A]. Elsevier: Amsterdam, 2007: 979-1030.

[29] 丁树良,汪文义,杨淑群. 认知诊断测验蓝图的设计 [J]. 心理科学,2011,34(2):258-265.

[30] 丁树良,杨淑群,汪文义. 可达矩阵在认知诊断测验编制中的重要作用 [J]. 江西师范大学学报: 自然科学版,2010,34(5):490-495.

[31] Liu Jingchen, Xu Gongjun, Ying Zhiliang. Data-driven learning of *Q*-matrix [J]. Applied Psychological Measurement, 2012, 36(7): 548-564.

[32] 喻晓锋,罗照盛,秦春影,等. 基于作答数据的模型参数和 *Q* 矩阵联合估计 [J]. 心理学报,2015,47(2):273-282.

The *Q*-Matrix Validation Methods and Comparison to Three Existing Methods

SONG Lihong[1], WANG Wenyi[2], DING Shuliang[2]

(1. Elementary Educational College, Jiangxi Normal University, Nanchang Jiangxi 330022, China;
2. College of Computer Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: The *Q*-matrix plays an important role in establishing the relation between latent attribute patterns and ideal response patterns. In practice, the *Q*-matrix is difficult to specify correctly in cognitive diagnostic assessment and misspecification of the *Q*-matrix can seriously affect the accuracy of both item parameter estimates and the classification of examinees. In the study, three on-line calibration methods have been extended to validate *Q*-matrix, and three related methods including the δ method, the γ method, and the *Q*-matrix refinement method (denoted by RSS) have been compared. A simulation study was conducted to investigate the sensitivity of validation methods to four factors (the distribution of attribute patterns, sample size, the quality of items, and the error rate of q-entries) under the deterministic inputs, noisy "and" gate (DINA) model. Results show that marginal maximum likelihood method performs best, both in terms of accuracy and robustness.

Key words: cognitive diagnostic assessment; the *Q*-matrix; the DINA model; the EM algorithm; on-line calibration method

(责任编辑: 冉小晓)