心理科学进展 2014, Vol. 22, No. 6, 1036–1046
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2014.01036

项目反应理论中潜在心理特质“填补”的参数估计方法及其演变*

田 伟[1] 辛 涛[2] 康春花[3]

([1]北京师范大学教育学部; [2]北京师范大学发展心理研究所, 北京 100875)
([3]浙江师范大学教育学院, 金华 321004)

摘 要 在心理与教育测量中, 项目反应理论(Item Response Theory, IRT)模型的参数估计方法是理论研究与实践应用的基本工具。最近, 由于IRT模型的不断扩展与EM (expectation-maximization)算法自身的固有问题, 参数估计方法的改进与发展显得尤为重要。这里介绍了IRT模型中边际极大似然估计的发展, 提出了它的阶段性特征, 即联合极大似然估计阶段、确定性潜在心理特质“填补”阶段、随机潜在心理特质“填补”阶段, 重点阐述了它的潜在心理特质“填补” (data augmentation)思想。EM算法与Metropolis-Hastings Robbins-Monro (MH-RM)算法作为不同的潜在心理特质“填补”方法, 都是边际极大似然估计的思想跨越。目前, 潜在心理特质“填补”的参数估计方法仍在不断发展与完善。

关键词 项目反应理论; 潜在心理特质; “填补”; 边际极大似然函数估计; EM算法; MH-RM算法
分类号 B841

1 引言

一般来说, 在心理与教育测量中, 经过严格的流程编制出心理或教育测验之后, 再通过经典测验理论或项目反应理论对被试的潜在心理特质(例如, 认知、人格、学业能力等)进行量尺化。其中, 项目反应理论因为具有相对于经典测验理论的一些优势(Brennan, 2006), 在实践中得到了广泛应用。无论是在大规模测验还是理论研究中, 项目反应理论都越来越成为主导的测验理论。

其中, IRT模型的参数估计方法是IRT理论研究及其实践应用的基本工具。Bock和Lieberman (1970)的边际极大似然估计 (maximum marginal likelihood estimation, MMLE) 是参数估计理论的基础, 但是它的计算速度很慢。到了Bock和Aitkin (1981), MMLE的计算速度才借以EM算法(Dempster, Laird, & Rubin, 1977)解决。最近, 由于IRT模型的不断扩展与EM算法本身的固有问题, 改进EM算法与开发更加高效准确的新算法成为IRT模型参数估计领域的研究热点。

IRT模型应用范围的不断扩展带来两个问题。首先, 不同的应用情景可能需要对单维0-1评分IRT模型进行扩展。例如, 多级评分IRT模型、多维IRT模型或者多维多水平随机效应IRT模型都是其扩展; 其次, IRT模型的扩展使得EM算法的固有问题显现, 即潜在心理特质正态分布假设不满足与“高维积分困境” (Wirth & Edwards, 2007)。针对这两个问题, 新的参数估计理论开始推动IRT模型软件开发与实践应用的不断发展。目前, 与经典的IRT模型参数估计软件(例如, BILOG, Zimowski, Muraki, Mislevy, & Bock, 2003; MULTILOG, Thissen, 2003; PARSCALE, Muraki & Bock, 1997)相比而言, 新的IRT模型参数估计软件(例如, IRTPRO, Cai, Thissen, & du Toit, 2011; flexMIRT, Cai, 2012a)在算法实施上有了很大的变化, 同时还在不断改进。

Baker (1992)与Baker和Kim (2004)的《项目

收稿日期: 2012-11-13
* 中央高校基本科研业务费专项资金资助(SKZZX2013028)、国家自然科学基金(31371047)、浙江省哲学社会科学规划基金(10CJY15YBB)、浙江省教育厅课题(Y2010117786)支持。
通讯作者: 辛涛, E-mail: xintao@bnu.edu.cn

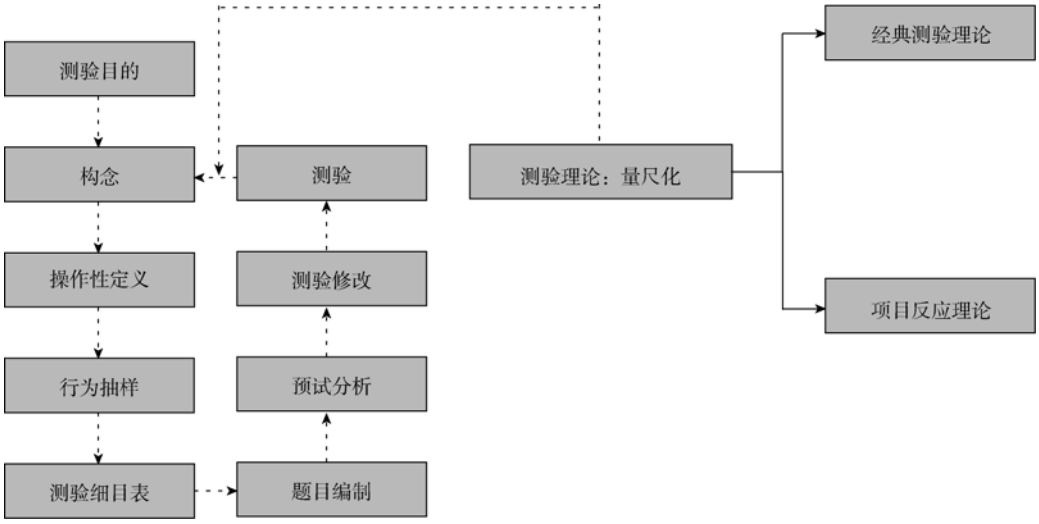

图 1　心理与教育测验及其统计理论

反应理论：参数估计理论与技术》是 IRT 模型参数估计领域最著名的理论专著。从 2004 年的再版可见，作者列出一章专门介绍马尔科夫链蒙特卡洛积分(Markov Chain Monte Carlo, MCMC)。MCMC 参数估计(Patz & Junker, 1999, Wirth & Edwards, 2007)是自 Albert (1992)以来，IRT 模型参数估计领域的进展之一。除此之外，漆书青、戴海琦和丁树良(1998)与 de Ayala (2009)同样介绍了 IRT 模型的参数估计方法。从 Baker (1992)到 de Ayala (2009)都在讨论 MMLE/EM 算法，但是都没有阐述它的统计思想—— 潜在心理特质“填补”。对统计思想的重视不足虽然不致引起参数估计偏差，但是不便于理解题目参数误差协方差矩阵的计算偏差。除此之外，不阐述统计思想也很难从整体上掌握 IRT 模型的参数估计框架。例如, EM 算法与 MH-RM 算法(Cai, 2010a, 2010b)分别是 MMLE 框架下的确定性潜在心理特质“填补”与随机潜在心理特质“填补”。最重要的是，MMLE/EM 算法在持续发展中已经融合了新的理论技术(例如, Bock & Moustaki, 2007; Cai, 2010c; Cai, Yang, & Hansen, 2011; Woods & Thissen, 2006)，同时也在影响着最新的 MH-RM 算法(Monroe & Cai, 2012)。因此，参数估计理论作为 IRT 理论发展与实践应用的关键组成部分，有必要阐述它的统计思想及其发展演变。

鉴于此，本研究将主要介绍 MMLE 的潜在心理特质“填补”方法及其演变。首先，将介绍 IRT 模型中的潜在心理特质“填补”思想；其次，将 IRT 模型的参数估计发展历史分为三个不同阶段，即联合极大似然估计阶段、确定性潜在心理特质“填补”阶段与随机潜在心理特质“填补”阶段；最后，提出应用这些参数估计方法的问题与研究展望。

2　潜在心理特质“填补”的思想

广义上，IRT 模型中的潜在心理特质可以看作缺失数据(missing data)。如果“填补”潜在心理特质，那么原先复杂的 MMLE 的最大化问题就转换成了一系列简单的完整数据对数似然函数最大化问题。因此，只要给定题目参数值初值和观测数据，潜在心理特质很容易被“填补”，“填补”后又容易最大化，这样就可以使用潜在心理特质“填补”的思想进行 IRT 模型参数估计。随着 IRT 模型的不断发展，潜在心理特质“填补”思想将持续发挥应有价值。

2.1　IRT 模型

假定一个测验由 $j=1,\cdots,n$个题目组成，考察了 $p\geq 1$个潜在心理特质。对于题目 j 来说，有 $K_j=\{0,1,\cdots,K_j-1\}$个反应类别，题目 j 的参数记为$\boldsymbol{\gamma}_j$。进而，n个题目的参数可以表示为$\boldsymbol{\gamma}$。假定被试的潜在心理特质可以用一个 p 维向量$\mathbf{x}$来刻画，那么$i=1,\cdots,N$个被试的潜在心理特质就可以记为一个$N\times p$的矩阵$\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_N)'$。最后，$N$ 个被试在 n 个题目上的作答反应矩阵记为：

$$\mathbf{Y}=\begin{pmatrix}\mathbf{y}_1\\ \vdots\\ \mathbf{y}_i\\ \vdots\\ \mathbf{y}_N\end{pmatrix}=\begin{pmatrix}y_{11} & \cdots & y_{1j} & \cdots & y_{1n}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ y_{i1} & \cdots & y_{ij} & \cdots & y_{in}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ y_{N1} & \cdots & y_{Nj} & \cdots & y_{Nn}\end{pmatrix}$$

一般地，潜在心理特质为$\mathbf{x}_i$的被试在题目j上作答反应为k的概率P_{ijk}可以表示为：

$$P_{ijk}(\mathbf{x})\triangleq P(y_{ij}=k|\mathbf{x}_i,\boldsymbol{\gamma}_j). \tag{1}$$

公式(1)是常见 IRT 模型的一般表示。例如，常见的单维与多维的 1/2/3 参数 logistic 模型、等级反应模型或者称名反应模型等(例如，Baker & Kim, 2004; Cai et al., 2011; Reckase, 2009)。对于等级反应模型来说，假定$\boldsymbol{\beta}_j=(\beta_{1,j},\cdots,\beta_{K_j-1,j})$是题目$j$的$(K_j-1)$维截距向量，$\boldsymbol{\alpha}_j=(\alpha_{1j},\cdots,\alpha_{pj})$是$p$维斜率参数向量，$P_+(k|\mathbf{x}_i,\boldsymbol{\gamma}_j)$是得分在$k$分及以上的概率，于是，每一个反应类别的累积条件反应概率(conditional cumulative response probability)可以表示为：

$$P_+(1|\mathbf{x}_i,\boldsymbol{\gamma}_j)=\frac{1}{1+\exp(-\beta_{j,1}-\boldsymbol{\alpha}_j'\mathbf{x}_i)}$$

$$\vdots$$

$$P_+(k|\mathbf{x}_i,\boldsymbol{\gamma}_j)=\frac{1}{1+\exp(-\beta_{j,k}-\boldsymbol{\alpha}_j'\mathbf{x}_i)}$$

$$\vdots$$

$$P_+(K_j-1|\mathbf{x}_i,\boldsymbol{\gamma}_j)=\frac{1}{1+\exp(-\beta_{j,K_j-1}-\boldsymbol{\alpha}_j'\mathbf{x}_i)}$$

如果规定$P_+(0|\mathbf{x}_i,\boldsymbol{\gamma}_j)=1$ 与$P_+(K_j|\mathbf{x}_i,\boldsymbol{\gamma}_j)=0$，那么，类别反应概率(category response probability)是两个相邻累积反应概率之差：

$$P(k|\mathbf{x}_i,\boldsymbol{\gamma}_j)=P_+(k|\mathbf{x}_i,\boldsymbol{\gamma}_j)-P_+(k+1|\mathbf{x}_i,\boldsymbol{\gamma}_j)$$

当$K_j=2$时，等级反应模型就变成了常见的多维 2PL 模型。这里的题目参数$\boldsymbol{\gamma}_j=(\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)$。

2.2 观测数据的边际似然函数

在实际应用中，潜在心理特质$\mathbf{x}_1,\cdots,\mathbf{x}_N$与题目参数$\boldsymbol{\gamma}$一般未知。Bock 和 Lieberman (1970)的 MMLE 是 IRT 模型参数估计理论与软件开发的经典范式。EM 算法(例如, Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988)或是 MH-RM 算法，都是先估计题目参数(calibration)再估计被试潜在心理特质(scoring) (例如，Cai, 2012a; Cai et al., 2011)。假定被试i在题目j上反应$y_{ij}=k$的指示函数(indicator function)为：

$$\chi_k(y_{ij})=\begin{cases}1, & y_{ij}=k,\\ 0, & \end{cases} \tag{2}$$

一般地，被试i在题目j上反应为y_{ij}的条件概率可以表示为：

$$P(y_{ij}|\mathbf{x}_i,\boldsymbol{\gamma}_j)=\prod_{k=0}^{K_j-1}P_{ijk}^{\chi_k(y_{ij})}. \tag{3}$$

进而，根据局部独立性假设(Lord & Novick, 1968)，作答反应矩阵$\mathbf{Y}$的边际似然函数可以表示为：

$$L(\boldsymbol{\gamma}|\mathbf{Y})=\prod_{i=1}^{N}\left[\int\prod_{j=1}^{n}P(y_{ij}|\mathbf{x}_i,\boldsymbol{\gamma}_j)\phi(\mathbf{x})d\mathbf{x}\right] \tag{4}$$

为了参数估计之便，假定潜在心理特质的总体分布$\phi(\mathbf{x})$为标准正态分布(例如，Lord & Novick, 1968; Bock & Aitkin, 1981; Bock et al., 1988; Cai, 2010a)。IRT 模型参数估计是要最大化$\log L(\boldsymbol{\gamma}|\mathbf{Y})$以得到题目参数的极大似然估计值$\hat{\boldsymbol{\gamma}}$。事实证明，直接最大化$\log(\boldsymbol{\gamma}|\mathbf{Y})$不具实践价值(Bock & Lieberman, 1970; Bock & Aitkin, 1981)，而是最大化完整数据对数似然函数的条件期望。

2.3 完整数据的对数似然函数

潜在心理特质可以看作$N\times p$的缺失数据矩阵$\mathbf{X}$，“填补”$\mathbf{X}$，得到完整数据矩阵$\mathbf{Z}=(\mathbf{X},\mathbf{Y})$，即：

$$\mathbf{Z}=\left(\begin{array}{c|ccccc}\mathbf{x}_1 & y_{11} & \cdots & y_{1j} & \cdots & y_{1n}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \mathbf{x}_i & y_{i1} & \cdots & y_{ij} & \cdots & y_{in}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \mathbf{x}_N & y_{N1} & \cdots & y_{Nj} & \cdots & y_{Nn}\end{array}\right)$$

于是，完整数据矩阵$\mathbf{Z}$的似然函数可以表示为：

$$L(\boldsymbol{\gamma}|\mathbf{Z})=\left[\prod_{i=1}^{N}\phi(\mathbf{x}_i)\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{n}P(y_{ij}|\mathbf{x}_i,\boldsymbol{\gamma}_j)\right] \tag{5}$$

公式(5)与公式(4)相比而言，更为简单、易于计算。而且，通过完整数据$\mathbf{Z}$的对数似然函数

$$\log L(\boldsymbol{\gamma}|\mathbf{Z})=C+\sum_{j=1}^{n}\left[\sum_{k=0}^{K_j-1}\sum_{i=1}^{N}\chi_k(y_{ij})\log P_{ijk}\right] \tag{6}$$

可以看到，题目参数可以逐题估计。因此，可以通过公式(5)解决公式(4)的最大化问题。

2.4 迭代的潜在心理特质“填补”

潜在心理特质“填补”的参数估计方法就是从后验分布$F(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^*)$中(随机)抽样来“填补”潜在心理特质$\mathbf{X}$，再最大化$\log L(\boldsymbol{\gamma}|\mathbf{Z})$的条件期望：

$$\mathcal{Q}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)=\int_{\varepsilon}\log L(\boldsymbol{\gamma}|\mathbf{Z})F(d\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^*). \tag{7}$$

这里$\mathbf{X}\in\varepsilon$，ε是一个样本空间。$F(d\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^*)$依赖于题目参数的当前估计值$\boldsymbol{\gamma}^*$。由于参数估计是从一

组给定的题目参数初值开始, 因此开始阶段估计得到的$\boldsymbol{\gamma}^*$是不准确的。从而, 后验分布$F(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^*)$与$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)$也是不准确的, 但是参数估计到了最后阶段, $F(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^*)$与$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)$会逐渐趋于稳定。那时, 最大化$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)$就可以得到题目参数的极大似然估计值$\hat{\boldsymbol{\gamma}}$。这就是潜在心理特质“填补”的参数估计方法, 将不完整数据估计问题转换为一系列迭代的完整数据估计问题(例如, Bock & Aitkin, 1981; Bock et al., 1988; Cai, 2010a, 2010b, 2010c; Cai et al., 2011)。

3　潜在心理特质“填补”方法的演变

由于测验实践中题目参数与被试潜在心理特质参数都是未知的, 因此这里讨论的 IRT 模型参数估计的历史发展限定从联合极大似然估计(joint maximum likelihood estimation, JMLE) (Lord & Novick, 1968)开始。根据 IRT 模型参数估计的任务与使用的统计方法将其历史发展分为三个阶段, 即联合极大似然估计阶段、确定性潜在心理特质“填补”阶段和随机潜在心理特质“填补”阶段。

第一阶段, 从 JMLE (Lord & Novick, 1968)到 Bock 和 Lieberman (1970)的 MMLE 算法。这一阶段的主要任务是解决题目参数估计的准确性问题。因为 JMLE 交替迭代地联合估计题目参数与被试潜在心理特质参数, 使得题目参数估计值不是一致估计量。理论上, Bock 和 Lieberman (1970)的 MMLE 方法解决了 JMLE 的一致性估计量问题。但是, 直到 20 世纪 80 年代左右研究者才认识到 JMLE 得到的估计是有偏的(Lord, 1983, 1986)。

第二阶段, 从 MMLE/EM 算法(Bock & Aitkin, 1981)的产生到 MMLE/EM 算法维度化简技术的完整扩展(Cai et al., 2011)。这一阶段主要是针对 EM 算法在 IRT 模型扩展中遇到的问题进行改进完善。最初, Bock 和 Aitkin (1981)应用统计上的 EM 算法(Dempster et al., 1977)显著提高了 MMLE 的参数估计效率, 从而 MMLE/EM 算法成为主流的参数估计方法。最近, 随着 IRT 模型在心理学中的广泛应用与多维 IRT 模型的流行, 潜在心理特质的正态分布假设受到质疑, “高维积分困境” (Wirth & Edwards, 2007)也变成难题。因此, Ramsay 曲线-IRT 与维度化简(dimension reduction)技术引入以解决参数估计的准确性与效率问题。这一阶段的共同特点是使用确定性的潜在心理特质“填补”。

第三阶段, 是随机潜在心理特质“填补”的产生与发展, 以解决“高维积分困境”问题为主要任务。将 MMLE 与 MCMC 结合(MMLE/MCMC), 可以抽取p维潜在心理特质向量。这一阶段的代表性算法是 MH-RM 算法, 它已经在 IRTPRO (Cai et al., 2011)中实施, 促进了它在实践中的广泛应用(例如, Cai, 2010a, 2010b; von Davier & Sinharay, 2010)。

3.1　联合极大似然估计阶段

当时, Bock 和 Lieberman (1970)MMLE 方法的理论价值远胜于实践价值, 因此 20 世纪 70 年代到 80 年代的 IRT 软件都使用 JMLE 方法。当$p=1$估计题目参数$\boldsymbol{\gamma}$与被试潜在心理特质参数$\mathrm{x}_1,\cdots,\mathrm{x}_N$时, 对于 JMLE 来说, 一次迭代由两个阶段组成：(1)计算N个被试原始分数的标准分数, 视标准分数$\mathrm{x}_1,\cdots,\mathrm{x}_N$为已知, 逐题估计题目参数$\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_n$; (2)假定$\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_n$是已知的, 估计$N$个被试的潜在心理特质。从 JMLE 的迭代过程可以看到, 给出$\mathrm{x}_1,\cdots,\mathrm{x}_N$后再估计题目参数$\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_n$使得参数估计问题大大简化。但是, 由于最初的$\mathrm{x}_1,\cdots,\mathrm{x}_N$是不准确的, 所以$\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_n$的估计值也是不准确的。因此, 不断地迭代使得$\mathrm{x}_1,\cdots,\mathrm{x}_N$越来越准确, 从而使得$\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_n$越来越准确。JMLE 与潜在心理特质“填补”方法类似, 第一阶段给出$\mathrm{x}_1,\cdots,\mathrm{x}_N$, 使得参数估计问题简化, 只是给的方式不同。第二个阶段最大化完整数据$\mathbf{Z}$的对数似然函数或者$\mathbf{Z}$的对数似然函数的期望。可见, JMLE 已经初步显现出潜在心理特质“填补”方法的思想萌芽。

3.2　确定性潜在心理特质“填补”阶段

在统计学中, Dempster 等人(1977)提出了 EM 算法, 不仅完整介绍了 EM 算法的理论基础, 而且还列举了 EM 算法的 7 种广泛应用。对于 IRT 模型的参数估计来说, 这篇文章传达出了潜在心理特质“填补”的统计思想。受到 Dempster 等人(1977) EM 算法的影响, Bock 和 Aitkin (1981)对 MMLE 重新作了复杂的求导与积分运算, 得到了经典的 Bock-Aitkin EM 算法, 标志着 IRT 模型中确定性潜在心理特质“填补”的提出。最近, 由于潜在心理特质的正态分布假设与“高维积分困境”问题, 出现了 Ramsay 曲线-IRT 与维度化简的 EM 算

法, EM 算法自身的固有问题也得以改善。为了更好地体现潜在心理特质“填补”思想，下面将撇开 Bock 和 Aitkin (1981)的冗长推导，着重 EM 算法的期望(expectation)思想，这对于理解 EM 算法误差协方差矩阵(error covariance matrix)的计算非常重要。

3.2.1 确定性潜在心理特质“填补”的产生——Bock-Aitkin EM 算法

IRT 模型中的 EM 算法推导都是 Bock 和 Aitkin (1981)的思路(Bock & Aitkin, 1981; Muraki, 1992)，而不是从 EM 算法的期望思想出发。这里，将从潜在心理特质“填补”思想入手介绍 EM 算法在 IRT 模型中的应用，并将 0-1 评分与多级评分的 EM 算法通过指示函数$\chi_k(y_{ij})$综合到一个框架下。潜在心理特质“填补”思想的关键是计算公式(7)定义的完整数据对数似然函数的期望。为此，要从$F(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^*)$中抽取$\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_N)'$“填补”$\mathbf{Y}$。对于第$i$个被试的潜在心理特质$\mathbf{x}_i$来说，就是从后验分布$F(\mathbf{x}_i|\mathbf{y}_i,\boldsymbol{\gamma}^*)$中“填补”。当然，“填补”有确定性与随机之分，以下首先介绍确定性的潜在心理特质“填补”。根据贝叶斯定理，第i个被试潜在心理特质的后验分布可以表示为：

$$F(\mathbf{x}_i|\mathbf{y}_i,\boldsymbol{\gamma}^*)=\frac{L(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\gamma}^*)\phi(\mathbf{x}_i)}{\int L(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\gamma}^*)\phi(\mathbf{x})d\mathbf{x}} \tag{8}$$

当 $p=1$ 时，如果从 $F(\mathrm{x}_i|\mathbf{y}_i,\boldsymbol{\gamma}^*)$ 中抽取 $\mathrm{x}_1,\cdots,\mathrm{x}_Q$个潜在心理特质“填补”，那么就可以利用这些离散的积分节点首先近似地计算公式(4)中括号项的边际概率为：

$$\bar{L}_i=\sum_{q=1}^{Q}\prod_{j=1}^{n}[\prod_{k=0}^{K_j-1}P(y_{ij}=k|\mathrm{x}_q,\boldsymbol{\gamma}_j^*)^{\chi_k(y_{ij})}]W_q \tag{9}$$

其中，$W_q=\frac{\phi(\mathrm{x}_q)}{\sum_{q=1}^{Q}\phi(\mathrm{x}_q)}$。从而，可以计算得到后验分布$F(\mathrm{x}_i|\mathbf{y}_i,\boldsymbol{\gamma}^*)$。最后，公式(7)可以表示为：

$$\mathcal{Q}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)=\sum_{j=1}^{n}\sum_{q=1}^{Q}\sum_{k=0}^{K_j-1}\bar{r}_{jk}(\mathrm{x}_q)\log P(y_{ij}=k|\mathrm{x}_q,\boldsymbol{\gamma}_j^*) \tag{10}$$

之所以称之为确定性的潜在心理特质“填补”，是因为$\mathrm{x}_1,\cdots,\mathrm{x}_Q$在迭代过程中固定不变。其中，$\bar{r}_{jk}(\mathrm{x}_q)$是基于作答反应矩阵$\mathbf{Y}$和当前题目参数估计值$\boldsymbol{\gamma}^*$计算得到的第$q$个积分节点在第$k$个反应类别上的“期望作答人数”，即：

$$\bar{r}_{jk}(\mathrm{x}_q)=\sum_{i=1}^{N}\chi_k(y_{ij})\frac{L_i(\mathrm{x}_q,\boldsymbol{\gamma}^*)W_q}{\bar{L}_i} \tag{11}$$

M 步根据牛顿—拉夫逊方法最大化公式(10)，迭代更新参数估计值为$\boldsymbol{\gamma}^{(*+1)}$。

3.2.2 基于 Ramsay 曲线的潜在心理特质“填补”

潜在心理特质分布$\phi(\mathbf{x})$的正态分布假设是现代 IRT 模型参数估计理论的基本假设(例如, Bock & Moustaki, 2007; Cai, 2010a)。当$p=1$时，如果正态分布假设成立，在近似潜在心理特质的后验分布$F(\mathrm{x}_i|\mathbf{y}_i,\boldsymbol{\gamma}^*)$时，权重$W_q$可以根据正态分布的概率密度函数$\phi(\mathrm{x})$得到，统计软件都有标准语句得到$\phi(\mathrm{x}_q),q=1,\cdots,Q$，从而可以很方便地计算$W_q$。然而，对于一些心理变量，正态分布假设是不合理的(例如，Woods, 2006; Woods & Thissen, 2006)。在 EM 算法中，不准确的W_q则会影响公式(11) $\bar{r}_{jk}(\mathrm{x}_q)$的计算，从而得到不准确的 IRT 模型参数估计值。事实上，$\phi(\mathrm{x})$的正态分布假设的影响是多方面的，包括计算机自适应化考试的选题、测验的编制、模型与数据拟合、测验维度检验等。

如果不对$\phi(\mathrm{x})$强加正态分布假设，那么就要从观测数据自由估计$\phi(\mathrm{x})$。最近, Woods 和 Thissen (2006)提出了一种新的方法估计$\phi(\mathrm{x})$，即 Ramsay 曲线 IRT (RC-IRT)。由于公式(11) $\bar{r}_{jk}(\mathrm{x}_q)$中的权重$W_q$是通过 Ramsay 曲线估计得到，因此称之为基于 Ramsay 曲线的确定性潜在心理特质“填补”。基于 Ramsay 曲线的 EM 算法是要提高 IRT 模型参数估计的准确性，不致产生有偏的参数估计值而影响 IRT 模型的正确应用。RC-IRT 是在 EM 算法中融合了 Ramsay (2000)的样条密度(spline density)估计方法，M 步估计题目参数的同时，通过似然函数：

$$L_\phi\propto\prod_{q=1}^{Q}\phi(\mathrm{x}_q)^{\bar{r}_j(\mathrm{x}_q)} \tag{12}$$

从观测数据估计潜在心理特质分布$\phi(\mathrm{x})$，$\phi(\mathrm{x}_q)$表示第q个积分节点上的概率密度。这里，$\bar{r}_j(\mathrm{x}_q)=\sum_{k=0}^{K_j-1}\bar{r}_{jk}(\mathrm{x}_q)$是积分节点$\mathrm{x}_q$上的“期望被试人数”。基于当前的$\phi(\mathrm{x}_q),q=1,\cdots,Q$，可以估计得到权重$W_1^*,\cdots,W_q^*,\cdots,W_Q^*$。下一个 E 步，使用$W_1^*,\cdots,W_q^*,\cdots,W_Q^*$计算$\bar{r}_{jk}^{*+1}(\mathrm{x}_q)$，再用$\bar{r}_{jk}^{*+1}(\mathrm{x}_q)$最大化公式(12)，如此迭代往复，直到收敛标准满足。如果$\phi(\mathrm{x})$是正态分布或者近似正态分布，那么 RC-IRT 估计得到的$W_1^*,\cdots,W_q^*,\cdots,W_Q^*$将与正态分布计算相差不多。如果$\phi(\mathrm{x})$不是正态分布，RC-IRT 则可以校正题目参数与期望后验分数(expected a posteriori score, EAP)的估计偏差。Woods 和 Thissen (2006)的 RC-IRT 还影响了随机潜在心理特质“填补”，即 MH-RM 算法的改进(参

见 Monroe & Cai, 2012)。

目前, RC-IRT 只限于单维 IRT 模型的参数估计。当 $p=1$时, 取出积分节点 $\mathrm{x}_1,\cdots,\mathrm{x}_Q$, 潜在心理特质分布 $\phi(\mathrm{x}_q)$可以用 Ramsay 曲线表示为：

$$\phi(\mathrm{x}_q)=\frac{\exp[\int_{\mathrm{X}_1}^{\mathrm{X}_q} w(\mathrm{x})d\mathrm{x}]}{\int_{\mathrm{X}_1}^{\infty}\exp[\int_{\mathrm{X}_1}^{\mathrm{X}_t} w(\mathrm{x})d\mathrm{x}]dt} \tag{13}$$

这里, 每一个积分节点 $\mathrm{x}_q, q=1,\cdots,Q$上, 样条(spline) $w(\mathbf{x}_q)$ 是 m 个 B-样条函数(B-spline function)：

$$B_s^d=\frac{\mathrm{x}_q-\kappa_s}{h}B_s^{d-1}(\mathrm{x}_q)+\frac{\kappa_{s+d+1}-\mathrm{x}_q}{h}B_{s+1}^{d-1}(\mathrm{x}_q) \tag{14}$$

的线性组合：

$$w(\mathbf{x}_q)=\mathbf{B}(\mathrm{x},\boldsymbol{\kappa},d)\mathbf{c} \tag{15}$$

这里, $m=(d+1)$+结点(knot)个数-2个样条函数, $\mathbf{B}(\mathrm{x},\boldsymbol{\kappa},d)$是一个$Q\times m$的 B-样条函数矩阵, $\mathbf{c}$就是*(12)*要估计的$m\times 1$参数向量, d表示度(degree), $\boldsymbol{\kappa}$表示潜在心理特质量尺上的一个结点, h表示结点之间的距离, s表示 B-样条开始的那个结点。有时为了更好地估计, 一般赋予$\mathbf{c}$一个m维的多变量正态分布, 通过贝叶斯最大后验方法(maximum a posteriori)进行估计。

3.2.3　维度化简的潜在心理特质“填补”

Bock 等(1988)将 EM 算法扩展到了多维测验。然而, 当 $p\geq 2$时, 公式(4)和(7)是 p 维积分。随着测验维度线性增加, 积分节点数量呈级数增加, 出现“高维积分困境”。对此, 两种不同思路, 一个是当固定每个维度的积分节点数量后, 要显著减少交叉积分节点数量。利用积分维度化简技术, 即利用测验维度结构特征将原来的高维积分降低为若干个不同的低维积分, 再进行积分运算; 再一个是抛掉确定性的潜在心理特质“填补”, 改用随机潜在心理特质“填补”, 直接抽取 p 维潜在心理特质向量 $\mathbf{x}_i=(\mathrm{x}_{i1},\cdots,\mathrm{x}_{ip})$ (Cai, 2010a, 2010b), 不再需要数以万计的交叉积分节点。

Gibbons 和 Hedeker (1992)于 Bock 等人(1988)不久, 在双因子模型(bifactor model)中使用 EM 算法时提出了维度化简技术。但是, 经过 15 年之多, Gibbons 等(2007)、Cai (2010c)与 Cai, Yang 和 Hansen (2011)才建立起维度化简 EM 算法的一般框架。双因子模型与两层模型(two-tier model)都是验证性 IRT (confirmatory IRT)模型, 研究者可以指定因子(factor)个数、因子之间的相关、因子与观测变量之间的因子模式(factor pattern)以及其他模型参数的先验假设(Cai, 2010b)。当一个测验的因子与观测变量之间的因子模式具有双因子或是两层结构时(例如, Cai, Yang, & Hansen, 2011; Cai, 2010c), EM 算法就可以借用维度化简方法显著减少积分节点数量, 使得参数估计更加高效稳定。

一般来说, 测验维度可以分为 p 维主维度 $\mathbf{x}$ (primary dimension) 与 S 维具体维度 (specific dimension) $\boldsymbol{\eta}$。主维度之间可以相关, 但是主维度与具体维度之间、具体维度与具体维度之间是正交的。任何一个题目至少考查一个主维度, 至多考查一个具体维度。这里, 基于一般的两层模型介绍维度化简技术, 因为当 $p=1$时, 两层模型变为双因子模型。因此, 当对边际似然函数*(4)*的中括号项组织后, 就可以得到：

$$L(\boldsymbol{\gamma}|\mathbf{y}_i)=\int_{\mathcal{R}^p}\int_{\mathcal{R}^S}\prod_{j=1}^{n}P(y_{ij}|\mathbf{x}_i,\eta_i,\boldsymbol{\gamma}_j)d\mathbf{x}_i d\eta_i=$$
$$\int_{\mathcal{R}^p}\prod_{s=1}^{S}\int_{\mathcal{R}}[\prod_{j\in\mathcal{I}_s}P(y_{ij}|\mathbf{x}_i,\eta_{is},\boldsymbol{\gamma}_j)\phi(\eta_{is})d\eta_i]\phi(\mathbf{x}_i)d\mathbf{x}_i \tag{16}$$

根据公式(16), 可以将测验题目分成不同子集 $\mathcal{I}_s$, $\mathcal{I}_s$中的题目只考查了主维度$\mathbf{x}$和具体维度η_s。最终, 原来$(p+S)$维积分化简成了$(p+1)$维积分, 显著减少了交叉积分节点, 节省了运算时间, 而且参数估计也更准确。

3.3　随机潜在心理特质“填补”阶段

对于多维 IRT 模型来说, EM 算法作为确定性的潜在心理特质“填补”, 只要各个维度上的积分节点数量确定了, 整个高维分布的潜在心理特质抽样也是确定的, 数量是不可控制的, “高维积分困境”是固有问题。虽然维度化简技术可以降低积分维度, 从而减少积分节点数量, 但是心理学研究作为探索性与验证性研究的科学过程, “高维积分困境”却是维度探索不可避免的。只有抛掉确定性的潜在心理特质“填补”, 改用随机潜在心理特质“填补”, 直接抽取 p 维潜在心理特质向量 $\mathbf{x}_i=(\mathrm{x}_{i1},\cdots,\mathrm{x}_{ip})$ (Cai, 2010a, 2010b), 才能彻底解决“高维积分困境”。

3.3.1　随机潜在心理特质“填补”思想

Meng 和 Schilling (1996)使用蒙特卡洛 EM (Monte Carlo EM)算法近似计算公式(7)的积分, 标志着随机潜在心理特质“填补”在 IRT 模型参数估计中开始应用。但是, 蒙特卡洛 EM 算法在实践中没有得到广泛应用。随着题目参数估计值接近$\hat{\boldsymbol{\gamma}}$, 随机抽取的潜在心理特质数量要显著增加,

大大降低了蒙特卡洛 EM 算法的收敛速度。而且，在下一次 EM 迭代时，前一个 E 步随机抽取的潜在心理特质将全部丢弃，再根据 M 步新的题目参数估计值重新抽取潜在心理特质近似公式(7)的积分，没有高效利用潜在心理特质样本(Cai, 2010a, 2010b)。但是，蒙特卡洛 EM 算法展现的随机潜在心理特质“填补”思想具有吸引力，它是直接随机抽取 p 维潜在心理特质向量$\mathbf{x}_i=(\mathrm{x}_{i1},\cdots,\mathrm{x}_{ip})$。因此，潜在心理特质的抽样数量完全可以通过马尔科夫链的性质来控制。因此，只有随机潜在心理特质“填补”可以避免“高维积分困境”。Cai (2010a)的 Metropolis-Hastings 抽样是一个简单的马尔科夫链，容易实施，在实践中得到了广泛应用。

潜在心理特质“填补”的参数估计方法必然是两阶段的参数估计。第一阶段，抽取潜在心理特质近似公式(7)；第二阶段，根据(7)找到极大似然估计值$\hat{\boldsymbol{\gamma}}$。两个阶段作为一个循环不断迭代，直到$\hat{\boldsymbol{\gamma}}$收敛。给定当前题目参数估计值$\boldsymbol{\gamma}^*$, (1)构造马尔科夫链从潜在心理特质的后验分布$F(\mathbf{X}|\boldsymbol{\gamma}^*,\mathbf{Y})$中随机抽取潜在心理特质$\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}$, Rubin (1987)称之为多重“填补” (multiple imputation)，近似(7)为：

$$\mathcal{Q}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)=\frac{1}{M}\sum_{m=1}^{M}\log L(\boldsymbol{\gamma}|\mathbf{Z}_m) \tag{17}$$

(2)根据牛顿—拉夫逊方法或者 Robbins-Monro 随机逼近(stochastic approximation, SA)更新参数估计值为$\boldsymbol{\gamma}^{(*+1)}$。

3.3.2 MH-RM 算法作为随机潜在特质“填补”

多维、多水平 IRT 模型正在成为新一代测量理论(Cai, 2012b)，因此 MH-RM 算法的产生与发展是必然的。MH-RM 算法非常适于 IRT 模型的软件开发，即使一个测验包含很多题目、测量很多潜在心理特质、有成千上万的被试参加，MH-RM 算法都可以很容易地实施。理论上，Cai (2010a) 的 MH-RM 算法是通过 Metropolis-Hastings (MH) 抽样随机“填补”潜在心理特质，再利用 Robbins-Monro 随机逼近算法来估计题目参数。最近, Monroe 和 Cai (2012)还把 RC-IRT 的方法扩展到了 MH-RM 算法。Yang 和 Cai (2012)又将 MH-RM 算法扩展到了多水平潜变量模型(multilevel latent variable modeling)估计背景效应(contextual effects)。

对于第i个被试，从后验分布$F(\mathbf{x}_i|\mathbf{y}_i,\boldsymbol{\gamma}^*)$中随机“填补”潜在心理特质$\mathbf{x}_i$。根据 Gibbs 方法从潜在心理特质的全条件分布(full conditional density)：

$$p(\mathbf{x}_i^l|\mathbf{x}_1^l,\cdots,\mathbf{x}_{i-1}^l,\mathbf{x}_{i+1}^{(l-1)},\cdots,\mathbf{x}_N^{(l-1)},\mathbf{Y},\boldsymbol{\gamma}^*) \tag{18}$$

直接抽样(l 表示第l次 MH 抽样)。但是很难从全条件分布(18)中直接抽样，再根据：

$$p(\mathbf{x}_i^l|\mathbf{x}_1^l,\cdots,\mathbf{x}_{i-1}^l,\mathbf{x}_{i+1}^{(l-1)},\cdots,\mathbf{x}_N^{(l-1)},\mathbf{Y},\boldsymbol{\gamma}^*)\propto$$

$$L(\boldsymbol{\gamma}|\mathbf{Z})=f(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\gamma}^*)\phi(\mathbf{x}_i)\left[\prod_{h\neq i}^{N}\phi(\mathbf{x}_h)\prod_{j=1}^{n}f(y_{ij}|\mathbf{x}_h,\boldsymbol{\gamma}_j^*)\right] \tag{19}$$

构建接受概率为：

$$R_{\mathbf{x}_i^*}=min\left\{\frac{f(\mathbf{y}_i|\mathbf{x}_i^*,\boldsymbol{\gamma}^*)\phi(\mathbf{x}_i^*)}{f(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\gamma}^*)\phi(\mathbf{x}_i)},1\right\} \tag{20}$$

的 MH 抽样。首先，抽取一个$N\times p$的矩阵 $\mathbf{E}$，每一行$\mathbf{e}_i$服从$\mathcal{N}_p(\mathbf{0},c^2\mathbf{I}_p)$，用户可以改变$c$得到不同的接受概率(acceptance ratio)。其次，计算潜在心理特质的备选抽样(proposal draws) $\mathbf{X}^*=\mathbf{X}+\mathbf{E}$。再次，计算每一个$\mathbf{x}_i^*$的接受概率$R_{\mathbf{x}_i^*}$以判断是接受还是拒绝。

对于第t次 MH-RM 迭代，通过 MH 抽样可以得到潜在心理特质$\{\mathbf{X}_l^t;l\geq 0\}$，去掉没有达到稳定之前的随机“填补” (burn-in period)，可以得到“填补”$\left\{\mathbf{X}_j^t;j=1,\cdots,m_t\right\}$近似计算公式(7)。对于第$(t+1)$次 MH-RM 迭代，MH 抽样的初始状态可以选作$\left\{\mathbf{X}_j^t;j=1,\cdots,m_t\right\}$的最后一个潜在心理特质抽样，即$\mathbf{X}_0^{(t+1)}=\mathbf{X}_{m_t}^t$。可见，MH-RM 算法不像蒙特卡洛EM那样将第$t$次迭代抽取的潜在心理特质抽样全部丢弃再重新抽样。一般来说，再高的测验维度 p，借助于统计软件(例如, MATLAB、R、SAS 等)都可以方便抽取$\mathbf{e}_i$。因此, MH 抽样从根本上解决了“高维积分困境”。从 IRTPRO 软件的使用情况来看，EM 算法需要几个小时的估计，MH-RM 算法只需几分钟到十几分钟就可完成。

3.4 小结：随机潜在心理特质“填补”成为必然趋势

一般来说，潜在心理特质是一个多维结构，不同潜在心理特质之间的关系有不同形式。而且，潜在心理特质会随着时空环境的变化而变化。为了准确刻画潜在心理特质，一般采用多维测验设计与分层随机抽样方案收集测验数据。因此，多维、多水平 IRT 模型正在成为新一代测量理论(Cai, 2012b)。然而，确定性潜在心理特质“填补”在参数估计时必然遇到“高维积分困境”。

Albert (1992)在 IRT 模型中引入了 MCMC,

从此 MCMC 在理论研究中不断发展(例如, Fox, 2003, 2005; Patz & Junker, 1999)。对于 MCMC 来说, 每一个被试的潜在心理特质与题目参数都是一个分布, 潜在心理特质与题目参数的联合后验分布$p(\mathbf{X},\boldsymbol{\gamma}|\mathbf{Y})$是抽取未知参数样本的目标分布(例如, Albert, 1992; Patz & Junker, 1999)。通过 Gibbs 抽样、MH 抽样或者 Gibbs 抽样与 MH 抽样的结合, 抽取随机参数样本 $M_0=(\mathbf{X}_0,\boldsymbol{\gamma}_0),M_1=(\mathbf{X}_1,\boldsymbol{\gamma}_1),M_2=(\mathbf{X}_2,\boldsymbol{\gamma}_2),\cdots$直至收敛到联合后验分布$p(\mathbf{X},\boldsymbol{\gamma}|\mathbf{Y})$为止。最后, 除去马尔科夫链没有达到稳定之前的随机样本, 其余的便可用于估计未知参数$\mathbf{X}$与$\boldsymbol{\gamma}$。

目前, 完全贝叶斯的 MCMC 方法只限在理论研究层面。一方面, 在心理与教育测量中, 没有参数估计软件可以实施 MCMC 抽样。即使借助于统计软件(例如, MATLAB、R、WINBUGS 等), 面对大规模教育测验与心理问卷调查也难以在实际中应用; 另一方面, IRT 模型作为统计模型, 模型与数据之间的拟合程度对于解释与揭示心理与教育现象本质是至关重要的。基于 MMLE 框架, 部分信息(limited information)拟合检验统计量在实际应用中发挥着重要作用(例如, Cai, 2008)。然而, 对于 MCMC 抽样方法, 尚未模型与数据之间的拟合检验统计量, 没有办法解释 IRT 模型对数据的代表性。因此, 当前参数估计理论的新进展是将完全贝叶斯统计的 MCMC 随机抽样与 MMLE 结合(MMLE/MCMC)。

4　问题与展望

目前, 确定性“填补”的 EM 算法与随机“填补”的 MH-RM 算法在实践中各有所用。例如, 对于单维 IRT 模型、p 不大的多维 IRT 模型或者验证性 IRT 模型(bifactor 与 two-tier)来说, 确定性潜在心理特质“填补”是高效准确的。对于 p 很大的探索性 IRT 模型或者多维、多水平 IRT 模型, 则不得不使用随机潜在心理特质“填补”。理论上, 对于 EM 算法与 MH-RM 算法来说, 都有需要扩展与改进之处。

4.1　EM 算法中题目参数误差协方差矩阵的计算理论与应用

心理与教育测量的目标是通过测验表现推测被试的潜在心理特质, 其准确性与测验工具的质量和参数估计方法有关。例如, 同一测验是否在不同施测组或者施测时间点具有测量不变性(例如, Lord, 1980; Bock, Muraki, & Pfeiffenberger, 1998; Cai et al., 2011)、IRT 模型与观测数据之间的拟合程度(例如, Cai, 2008; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Bartholomew & Leung, 2002)、题目参数估计值的不确定性等对被试潜在心理特质估计的影响(例如, Cheng & Yuan, 2010; Yang, Hansen, & Cai, 2012), 都需要计算题目参数的误差协方差矩阵。然而, Bock-Aitkin EM 算法在迭代终止时不能自然得到题目参数的误差协方差矩阵(Cai, 2008)。

4.1.1　题目参数误差协方差矩阵的数值计算

一直以来, EM 算法 M 步得到的完整数据协方差矩阵不能精确反映题目参数估计值的不确定性, 是因为它不含潜在心理特质作为缺失数据产生的不确定性(例如, Meng & Rubin, 1991; Cai, 2008)。

Cai (2008)将迭代的 Supplemented EM 算法(Meng & Rubin, 1991)用来计算题目参数的误差协方差矩阵, 但是 Supplemented EM 算法的计算时间可能比 EM 算法的计算时间还长(Cai, 2010b), 尤其是多维 IRT 模型。因此, 从理论上来看, 需要对 Supplemented EM 算法的计算效率、数值精确性进行改进。

4.1.2　认知诊断模型与题目参数误差协方差矩阵

认知诊断模型也多是基于 EM 算法估计题目参数与被试的知识状态。第一阶段估计题目参数, 第二阶段假定估计的题目参数为真值, 再估计被试的知识状态。然而, 由于抽样误差, 题目参数估计值并不完全等于真值。因此, 被试知识状态的估计没有考虑题目参数估计值的误差。在 IRT 模型中, 研究者刻画与校正了题目参数估计误差对潜在心理特质估计的影响(例如, Cheng & Yuan, 2010; Yang, Hansen, & Cai, 2012)。认知诊断模型中题目参数估计误差对估计被试知识状态的影响也是一个重要问题, 需要借助于题目参数的误差协方差矩阵。

4.2　潜在心理特质多变量正态分布假设

在心理与健康卫生领域, 当潜在心理特质非正态分布时, RC-IRT 在估计题目参数的同时估计潜在心理特质的概率分布(例如, Cai, Yang, & Hansen, 2011; Monroe & Cai, 2012; Woods, 2006; Woods & Thissen, 2006; Woods, 2007)。但是, RC-IRT 估计有诸多不便之处。例如, 概念复杂、

多个$\phi(\mathbf{x})$曲线的选择与解释、主流 IRT 软件还没有内置的 RC-IRT 估计等。

4.2.1 检验 **bifactor** 模型主维度的正态分布假设

为了避免复杂的 RC-IRT 估计，只有事前检验$\phi(\mathbf{x})$是否正态分布(Li & Cai, 2012)。如果潜在心理特质是正态分布，那么就用传统 EM 算法。如果不是正态分布，再采用 RC-IRT 估计。

Cai 和 Woods (2012)将经验直方图(Empirical Histogram)方法扩展到了双因子模型的主维度分布估计。对于双因子模型，不论使用 RC-IRT 还是经验直方图估计潜在心理特质的概率密度，都使本就繁重的参数估计更加耗时。因此，为了避免使用经验直方图估计概率密度，可以将 Li 和 Cai (2012)的方法扩展到双因子模型，构建一个基于测验总分的部分信息拟合统计量，检验主维度是否正态分布。

4.2.2 潜在心理特质多变量分布的估计

当测验考查的多维潜在心理特质不是正态分布时，RC-IRT 作为一般方法可以扩展到多变量分布估计。首先，在高维空间中，B-样条曲线(B-spline curve)变成 B-样条曲面(B-spline surface)，而且单维积分节点变成多维交叉积分节点。如何对公式(13)进行扩展需要进一步研究。其次，将 RC-IRT 推广到多维 IRT 不可避免地遇到“高维积分困境”。一方面，可以在 MH-RM 算法的框架下考虑多维 RC-IRT，另一方面，还可以将 RC-IRT 与双因子模型或者两层模型结合起来，通过维度化简方法降低参数估计难度。

4.3 基于模型整合的算法框架

当代心理与教育测量是信息挖掘的过程，是探索性与验证性分析结合的过程，是基于证据决策的过程。目前，实践中常用的模型包括 IRT 模型、多水平 IRT 模型、潜类别模型、混合 IRT 模型(mixture IRT)、认知诊断模型等，不同模型蕴含的信息各不相同。将这些模型整合到一个框架中，再基于整合的模型梳理算法结构，开发更为一般的心理与教育测量软件，更好地让测量理论与技术服务于实践，需要进一步研究。潜在心理特质“填补”的参数估计方法在一般模型框架下将显示出更大的应用空间，特别是对 MH-RM 算法的进一步扩展。

参考文献

漆树青，戴海琦，丁树良. (1998). *现代教育与心理测量学原理*. 北京：高等教育出版社.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251–269.

Baker, F. B. (Ed.). (1992). *Item response theory: Parameter estimation techniques*. New York: Dekker.

Baker, F. B., & Kim, S.-H. (Eds.). (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology, 55*, 1–15.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*, 179–197.

Bock, R. D., & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (vol. 26, pp. 469–513). The Netherlands: Elsevier B.V.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger Publishers.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309–329.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307–335.

Cai, L. (2010c). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581–612.

Cai, L. (2012a). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Cai, L. (2012b). *Second-generation, multidimensional, and multilevel item response modeling.* Paper presented at the National Council on Measurement in Education, Vancouver, BC.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2*p* tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173–194.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Lincolnwood, IL: Scientific Software International, Inc.

Cai, L., & Woods, C. (2012). *Maximum Marginal Likelihood Item Bifactor Analysis with Estimation of the General Dimension as an Empirical Histogram.* Paper presented at the 77th annual meeting of the psychometric society. Lincoln, Nebraska.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221–248.

Cheng, Y., & Yuan, K.-H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika, 75*, 280–291.

de Ayala, R. J. (Eds.). (2009). *The theory and practice of item response theory*. New York: Guilford.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society-Series B, 39*, 1–38.

Fox, J.-P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology, 56*, 65–81.

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology, 58*, 145–172.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K.,... Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4–19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

Li, Z., & Cai, L. (2012). *Summed score based fit indices for testing latent variable distribution assumption in IRT*. Paper presented at the 77th annual meeting of the psychometric society. Lincoln, Nebraska.

Lord, F. M. (Ed.). (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Lord, F. M. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika, 48*, 425–435.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157–162.

Meng, X. L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*, 899–909.

Meng, X. L., & Schilling, S. (1996). Fitting full-information factors models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association, 91*, 1254–1267.

Monroe, S., & Cai, L. (2012). *Estimation of a Ramsay-curve IRT model using the Metropolis-Hastings Robbins-Monro algorithm.* Paper presented at the 77th annual meeting of the psychometric society. Lincoln, Nebraska.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Muraki, E., & Bock, R. D. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data [Computer software]. Chicago: Scientific Software.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178.

Ramsay, J. O. (2000). Differential equation models for statistical functions. *Canadian Journal of Statistics, 28*, 225–240.

Reckase, M. D. (Ed.). (2009). *Multidimensional item response theory*. New York: Springer.

Rubin, D. B. (Ed.). (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Thissen, D. (2003). *MULTILOG 7 user's guide*. Chicago: SSI International.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics, 35*, 174–193.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.

Woods, C. M. (2006). Ramsay-Curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods, 11*, 253–270.

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement, 67*, 73–87.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*, 281–301.

Yang, J. S., & Cai, L. (2012). *Estimation of contextual effects through nonlinear multilevel latent variable*

modeling with a Metropolis-Hastings Robbins-Monro algorithm. Paper presented at the 77th annual meeting of the psychometric society. Lincoln, Nebraska.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement, 72*, 264–290.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG3 user's guide*. Chicago: SSI International.

The Data-augmentation Techniques in Item Response Modeling: Current Approaches and New Developments

TIAN Wei[1]; XIN Tao[2]; KANG Chunhua[3]

([1] *Faculty of Education*, *Beijing Normal University*, *Beijing* 100875, *China*)

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

([3] *College of Teacher Education*, *Zhejiang Normal University*, *Jinhua* 321004, *China*)

Abstract: The parameter estimation techniques in item response theory modeling are indispensable to theoretical researches and real applications. This paper focused on its data augmentation techniques and described its historical development from the Bock and Aitkin's (1981) deterministic EM algorithm to the Cai's (2010) Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (the integration of Markov Chain Monte Carlo and maximum marginal likelihood estimation, known as the stochastic data augmentation). Currently, the statistical computing still needs to be developed in new applications.

Key words: item response theory; latent trait; data augmentation; maximum marginal likelihood estimation; EM algorithm; MH-RM algorithm