

2PL 模型的两种马尔可夫蒙特卡洛缺失 数据处理方法比较*

曾莉¹ 辛涛² 张淑梅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北京师范大学心理学院, 北京 100875)

摘要 马尔可夫蒙特卡洛(MCMC)是项目反应理论中处理缺失数据的一种典型方法。文章通过模拟研究比较了在不同被试人数,项目数,缺失比例下两种 MCMC 方法(M-H within Gibbs 和 DA-T Gibbs)参数估计的精确性,并结合了实证研究。研究结果表明,两种方法是有差异的,项目参数估计均受被试人数影响很大,受缺失比例影响相对更小。在样本较大缺失比例较小时,M-H within Gibbs 参数估计的均方误差(RMSE)相对略小,随着样本数的减少或缺失比例的增加,DA-T Gibbs 方法逐渐优于 M-H within Gibbs 方法。

关键词 IRT; 缺失数据; MCMC。

分类号 B841.7

1 引言

项目反应理论(IRT)(漆书青,戴海琦,丁树良,2002)是教育与心理测量研究中的一种重要理论。在 IRT 的应用中,估计模型参数是一个重要的环节。在实际应用中,缺失数据的出现是很普遍的。不同的缺失数据处理方法会对参数估计产生影响(Karkee & Finkelman,2007;Ludlow & O'Leary,1999),因此,缺失数据处理方法的正确选择就变得尤为重要。

在项目反应有缺失时,删除含有缺失反应的被试是一种简单常用的方法,如 PARSCALE 软件(Muraki & Bock,1993)便是如此。但直接删除很可能会得到一个有偏的样本,从而得到有偏的估计,尤其在缺失比例很大的时候。且一个被试如果只是某些题目缺失,而不是所有题目全部缺失的时候,把整个被试的作答信息全部删除,我们就丢失了很多信息。因此,更好的方法不是删除被试,而是充分利用被试已作答的信息。

除了删除被试,目前 IRT 的缺失数据处理方法主要分为两类。一类是借补(Huisman,2000;Little & Rubin,2002/2004;Ludlow & O'Leary,1999),在参数估计前对数据进行预处理(用借补值替代缺失

值),把借补后的数据看作完整数据,再运用针对完整数据的参数估计方法。Ludlow 和 O'Leary(1999)基于 Rasch 模型,研究了对省略不答的试题和来不及回答的试题的不同的处理方法(即:均不用于参数估计;均看作不正确;省略不答的试题看作不正确,来不及回答的试题不用于参数估计;难度和能力估计的两步过程),得出的结论是不同的处理方法会对估计结果产生很大的影响。Huisman(2000)基于 Rasch 模型比较了一些简单的借补方法(替换为随机抽取值,不正确,项目均值,个体均值,正确项目均值,项目相关,热平台等),得出的结论是正确项目均值替代相对更好。另一类是缺失处理与模型相结合(Little & Rubin,2002/2004),应用到 IRT 的比较典型的是马尔可夫链蒙特卡洛(MCMC)方法(Karkee & Finkelman,2007;Maris & Bechger,2005;Patz & Junker,1999a;Ross,2006)。Karkee 和 Finkelman(2007)比较了四种缺失数据处理方法:把缺失反应视为错误;把缺失反应视为来不及回答;把最后五个项目的缺失视为来不及回答,其余的视为错误;MCMC 方法。得出的结论是把缺失反应视为来不及回答和 MCMC 两种方法相对更好。

MCMC 是一种多元分布的抽样方法,把它应用

收稿日期:2008-08-13

* 国家自然科学基金(30670718)项目资助。

通讯作者:辛涛,E-mail: xintao@bnu.edu.cn,电话:010-58805278

到 IRT 的参数估计会有很多优点(Patz & Junker, 1999a; Patz & Junker, 1999b): MCMC 是对参数的联合后验分布进行抽样, 因此它可以得到参数的联合后验分布的估计, 而不仅仅是参数的点估计; EM 算法(Little & Rubin, 2002/2004; 漆书青 等, 2002)需要精确的数值积分(E 步)及导数运算(M 步), 而 MCMC 方法则不需要, 因此 MCMC 方法更容易实现; 随着 IRT 模型变得更复杂, EM 算法的应用会越来越困难, 标准误的计算也变得更困难, 而 MCMC 估计方法能相对更容易地推广到更复杂的模型, 它的标准误计算方法是简单的; 更重要的是, 由于 MCMC 方法的灵活性, 它对缺失数据的处理是简单易行的(只需把缺失数据看作参数, 参与抽样即可)。

Albert (1992) 首先把一种特殊的 MCMC——Gibbs 抽样应用到了两参数正态卵状模型的参数估计。Patz 和 Junker (1999a; 1999b) 把另一种 MCMC——M-H within Gibbs 应用到了 2PL 模型的参数估计, 并把它推广到了更复杂的模型及处理缺失数据的情况。在 M-H within Gibbs 参数估计中, 需要给出建议密度(Patz & Junker, 1999a; Patz & Junker, 1999b), Jones 和 Nediak (2000) 在 Patz 和 Junker 的基础上给出了新的建议密度, 提高了有效性(链收敛得更快), 但没有涉及参数估计精度的提高。在 Albert 的基础上, Maris 和 Bechger (2005) 介绍了应用到 2PL 的参数估计的 DA-T Gibbs 方法, 但还没有把 DA-T Gibbs 应用到缺失处理。

目前对 MCMC 的讨论都是把 MCMC 与其他方法作比较, 但没有比较各种 MCMC 方法, 那不同的 MCMC 方法对缺失数据的参数估计是否会有差异, 这就需要我们进行进一步的研究。本文比较了应用到两参数逻辑斯蒂克(2PL) 0-1 评分模型(漆书青 等, 2002)的两种 MCMC 方法(M-H within Gibbs 和 DA-T Gibbs)对缺失数据的项目参数估计的影响。文章采用了模拟研究, 目的是比较在不同的被试人数, 项目数, 缺失比例下两种 MCMC 方法参数估计的精确性, 并且文章还进一步采用实证研究验证了模拟研究的结果。

2 模型

2.1 研究模型

2.1.1 两参数逻辑斯蒂克(2PL)模型 设项目数为 J , α_j 为第 j 题的区分度参数, β_j 为第 j 题的难度参数, $j = 1, 2, \dots, J$ 。记被试能力为 θ , $P_j(\theta)$ 为能力

为 θ 的被试在项目 j 上正确作答的概率, 则 2PL 模型(漆书青 等, 2002)为

$$P_j(\theta) = \frac{\exp[\alpha_j(\theta - \beta_j)]}{1 + \exp[\alpha_j(\theta - \beta_j)]}, j = 1, 2, \dots, J$$

在本文的研究中, 为了数学上的方便, 令 $\gamma_j = \alpha_j \times \beta_j$, 得到 2PL 模型的等价形式:

$$P_j(\theta) = \frac{\exp[\alpha_j(\theta - \gamma_j)]}{1 + \exp[\alpha_j(\theta - \gamma_j)]}, j = 1, 2, \dots, J$$

2.1.2 参数识别问题 2PL 模型的参数是不可识别的(Maris & Bechger, 2005), 因此需要对模型作出一定的假设。Tsutakawa (1985) 指出, 以下两个假设等价

$$(i) \prod_{j=1}^J \alpha_j = 1, \sum_{j=1}^J \beta_j = 0 (\beta_j = \gamma_j / \alpha_j),$$

$$\theta \sim N(\mu, \sigma^2)$$

$$(ii) \theta \sim N(0, 1)$$

这里我们选用第二个假设, 即不对项目参数作出限制, 而仅假设。

2.2 两种 MCMC 方法

基本的 MCMC 方法有 Metropolis-Hustings 算法, Gibbs 抽样等等(Ross, 2006), 其中 Gibbs 抽样是 M-H 算法的特例, 是一种广泛使用的 MCMC 方法。此外还有一些在这些基本方法基础上的扩展方法(Patz & Junker, 1999a)。IRT 模型参数个数很多, 所以对于大多数 IRT 模型直接应用 M-H 算法是不行的(Patz & Junker, 1999a)。Gibbs 抽样中涉及完全条件分布(Maris & Bechger, 2005; Patz & Junker, 1999a; Ross, 2006), 当完全条件分布易于处理时, Gibbs 抽样是有用的。但对于 2PL 模型, 完全条件分布是不易于处理的。针对 2PL 模型的这个问题, 目前有两个解决办法——M-H within Gibbs 抽样(Patz & Junker, 1999a; Patz & Junker, 1999b)和 DA-T Gibbs 抽样(Maris & Bechger, 2005)。

M-H within Gibbs 抽样是以 Gibbs 为基础, 对不易处理的完全条件分布再采用 M-H 抽样。即整体运用 Gibbs 抽样, 在抽样的过程中加入 M-H 算法, 从而实现目标后验分布的抽样。

DA-T Gibbs 抽样的思想是增加一个未知的潜在变量, 表面上看, 增加了未知变量, 需要估计的量增加, 但只要增加的变量合适, 往往会使 Gibbs 抽样中的完全条件分布变得易于处理。对于 2PL 模型, Maris 和 Bechger (2005) 引入一个服从 logistic 分布的变量后, 把此变量作为新的参数, 再运用 Gibbs 抽样, 此时所有参数的完全条件分布均是容易处理的, 从而实现了目标后验分布的抽样。

Patz 和 Junker (1999b) 指出,应用 M-H within Gibbs 方法进行参数估计时,如果反应矩阵有缺失,只需把缺失数据看作参数,与其他参数一起参与抽样即可。

对于 DA-T Gibbs 抽样,Maris 和 Bechger(2005) 在文中提到,处理缺失数据的方法是删除有缺失的被试,但前文已指出,直接删除是不好的。Little 和 Rubin(2002/2004) 认为,用 Gibbs 抽样处理缺失数据的时候,只需把缺失数据看作参数,和其余参数一起参与抽样即可。且 Patz 和 Junker(1999b) 运用 M-H within Gibbs 处理缺失数据的方法也是如此。因此,我们应用 DA-T Gibbs 时把缺失数据看作参数,参与 Gibbs 抽样过程。

3 模拟研究

3.1 研究设计

考虑到参数估计可能受到被试人数 N ,项目数 J ,缺失比例 p 等因素的影响,因此用蒙特卡洛方法产生被试的反应矩阵时我们考虑了两种项目数($J = 15, 40$),三种被试人数($N = 1000, 500, 100$),四种缺失比例($p = 0, 0.05, 0.1, 0.25$)。一共有 24 种组合情况。

取 $\alpha \sim \log \text{norm}(0, 0.5)$, $\gamma \sim N(0, 2)$, $\theta \sim N(0, 1)$ 。根据参数分布模拟产生出反应矩阵。再模拟缺失,对每个人每道题产生均匀分布随机数,若该随机数小于 p ,则认为该人在这道题上缺失。从而得到含有缺失反应的反应矩阵。此时我们认为缺失机制是完全随机缺失(Little & Rubin, 2002/2004)。

对于 J, N, p 的 24 种组合情况分别模拟含有缺失反应的反应矩阵,再分别运用两种 MCMC 方法估计项目参数,进而进行比较分析。其中,每种组合情

况均重复模拟 20 次。

3.2 比较标准

我们在 Patz 和 Junker(1999a) 给出的针对完整数据的 M-H within Gibbs 算法的 S-plus 程序的基础上运用 R software 编程实现了针对缺失数据的 M-H within Gibbs 及 DA-T Gibbs 参数估计方法。运用这两个程序进行参数估计,把估计的参数值与真实参数值作比较。

为衡量两个方法揭示真实参数的性能,运用指标 RMSE (root mean square errors) (漆书青,戴海琦,丁树良,2003):

$$\text{RMSE}(\alpha) = \sqrt{\frac{\sum_{j=1}^J \sum_{m=1}^M (\hat{\alpha}_j - \alpha_j)^2}{MJ}},$$

$$\text{RMSE}(\gamma) = \sqrt{\frac{\sum_{j=1}^J \sum_{m=1}^M (\hat{\gamma}_j - \gamma_j)^2}{MJ}}.$$

其中, M 为模拟重复次数, J 为项目数。RMSE 越小越好。

3.3 结果

对 24 套模拟反应矩阵分别运用两种 MCMC 方法进行参数估计(整个模拟及估计过程重复 20 次),计算出 RMSE,结果见表 1。

3.3.1 比较被试人数 N 对参数估计的影响 为了更直观地比较,我们画出 18 种组合情况的两种方法的 RMSE 散点图,如图 1 所示。其中横坐标“样本”表示被试人数 N 与缺失比例 p 的 12 种组合情况,具体样本代码见表 2(左),图中折线代表意义见表 2(右)。

图分为三个组,左边一组(样本 1, 2, 3) $p = 0.05$,中间一组(样本 4, 5, 6) $p = 0.1$,右边一组(样本 7, 8, 9) $p = 0.25$ 。

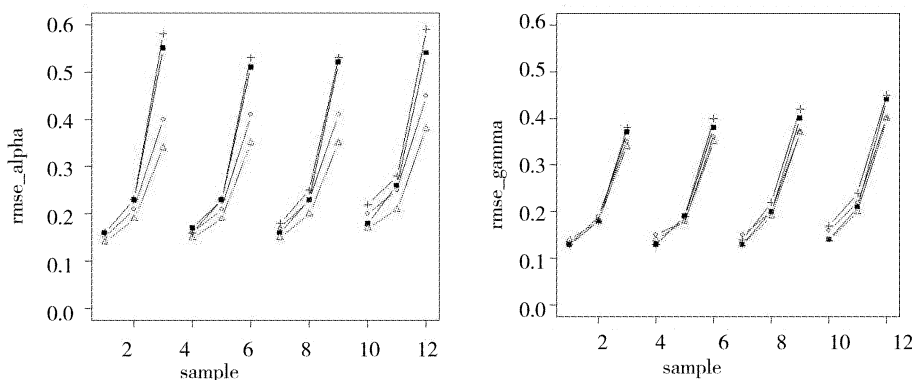


图1 参数 α 与 γ 的 RMSE 散点图

表 1 不同条件下两种方法得到的 RMSE

项目数 J	缺失比例 p	被试人数 N	RMSE			
			斜率参数 α		截距参数 γ	
			MH within Gibbs	DA-T Gibbs	MH within Gibbs	DA-T Gibbs
15	0	1000	0.16	0.15	0.13	0.13
		500	0.23	0.21	0.18	0.19
		100	0.58	0.40	0.38	0.35
	0.05	1000	0.16	0.16	0.13	0.15
		500	0.23	0.21	0.19	0.18
		100	0.53	0.41	0.40	0.36
	0.10	1000	0.18	0.17	0.14	0.15
		500	0.25	0.23	0.22	0.20
		100	0.53	0.41	0.42	0.37
	0.25	1000	0.22	0.20	0.17	0.16
		500	0.28	0.25	0.24	0.22
		100	0.59	0.45	0.45	0.40
40	0	1000	0.16	0.14	0.13	0.14
		500	0.23	0.19	0.18	0.18
		100	0.55	0.34	0.37	0.34
	0.05	1000	0.16	0.16	0.13	0.15
		500	0.23	0.21	0.19	0.18
		100	0.53	0.41	0.40	0.36
	0.10	1000	0.16	0.15	0.13	0.13
		500	0.23	0.20	0.20	0.19
		100	0.52	0.35	0.40	0.37
	0.25	1000	0.18	0.17	0.14	0.14
		500	0.26	0.21	0.21	0.20
		100	0.54	0.38	0.44	0.40

表 2 图 1 的注释(左为样本编号代码,右为折线符号代码)

被试人数 N	缺失比例 p				项目数 J	方法
	0	0.05	0.10	0.25		
1000	1	4	7	10	圆圈	15 DA-T Gibbs
500	2	5	8	11	三角形	40 DA-T Gibbs
100	3	6	9	12	加号	15 M-H within Gibbs
					方块	40 M-H within Gibbs

A. 当项目数为 15 时 为了比较两种方法的参数估计受被试人数 N 的影响,我们在图 1 的四个组中分别比较圆圈和加号折线。由观察可以看出,两种方法均受被试人数影响显著,随着被试人数的减少,参数估计均逐渐变差(RMSE 变大),尤其在中等样本(N = 500)变为小样本(N = 100)时,迅速变坏。

从大样本(N = 1000)变为中等样本时,两种方法的变坏速度差不多,但从中等样本变为小样本时,M-H within Gibbs 方法的变坏速度比 DA-T Gibbs 方法快得多,这说明 M-H within Gibbs 方法对被试人数更敏感。

首先看 p = 0 的情况(图中左边一组)。在大样

本时两种方法差异不大;中等样本时,DA-T Gibbs 方法略好;小样本时,DA-T Gibbs 方法明显优于 M-H within Gibbs 方法。

当 p = 0.05、0.1 和 0.25 时(图中中间两组和右边一组),参数估计的变化模式与 p = 0 时是类似的,只是随着缺失比例的增加,同一样本量下 DA-T Gibbs 方法优于 M-H within Gibbs 方法的优势有缓慢增加的趋势。

B. 当项目数为 40 时 此时在四个组中分别比较三角形和方块折线。观察可知,两种方法参数估计的变化模式与项目数为 15 时是类似的,只是对于 α 的估计两种方法的差异增大,对于 γ 的估计,两种方

法的差异略有减小。

3.3.2 比较缺失比例 p 对参数估计的影响 同样,我们画出 24 种组合情况的两种方法的 RMSE 散点

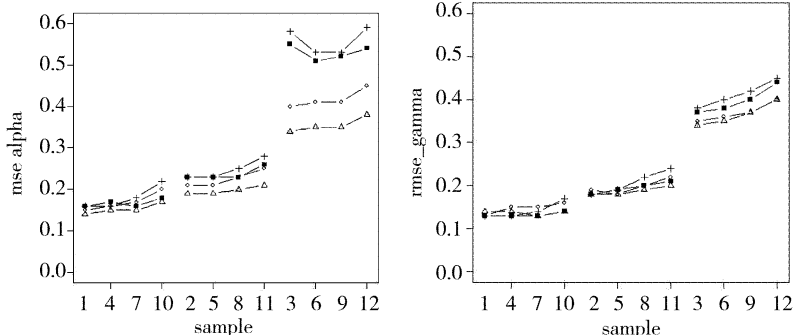


图2 参数 α 与 γ 的 RMSE 散点图

图分为三个组,左边一组(样本 1,4,7,10)的 $N = 1000$,中间一组(样本 2,5,8,11) $N = 500$,右边一组(样本 3,6,9,12) $N = 100$ 。

A. 当项目数为 15 时 为了比较两种方法的参数估计受缺失比例 p 的影响,我们在图 2 的三个组中分别比较圆圈和加号折线。由观察可以看出,随着缺失比例的增加,两种方法均有变坏的趋势,但当样本量很小时(右边一组),两种方法变得很不稳定。缺失比例从 0 变为 0.05 或从 0.05 变为 1 时,两种方法变坏的幅度相对较小,而缺失比例从 0.1 变为 0.25 时,变坏幅度相对更大。整体而言,两种方法均受缺失比例的影响是小于受样本量的影响的(在我们所研究的缺失比例范围内)。

大样本时(左边一组),对于 α 的估计,两种方法在缺失比例较小时是比较接近的,随着缺失比例的增加 DA-T Gibbs 变得更好;对于 γ 的估计,在缺失比例较小时, M-H within Gibbs 方法稍好,随着缺失比例的增加 DA-T Gibbs 变得更好。在中等样本时(中间一组),不管缺失比例是多少,均有 DA-T Gibbs 方法更好,且随着缺失比例的增加,两种方法之间的差异变得更大。在小样本时(右边一组),不管缺失比例是多少,均有 DA-T Gibbs 优于 M-H within Gibbs 很多,且随着缺失比例的增加,两种方法之间的差异变得更大。因此 M-H within Gibbs 受缺失比例的影响略大于 DA-T Gibbs 受缺失比例的影响,随着缺失比例的增加,DA-T Gibbs 方法优于 M-H within Gibbs 方法的优势有缓慢增加的趋势。

综合对样本量的影响的研究,我们有:在大样本小缺失的时候, M-H within Gibbs 略优于 DA-T Gibbs,但随着样本量的减少或缺失比例的增加,DA-T Gibbs 会反过来优于 M-H within Gibbs,且随着样

图,如图 2 所示。图中代码与图 1 相同,不同的是此时横坐标“样本”的顺序为:1,4,7,2,5,8,3,6,9。

本量的减少或缺失比例的增加,优势变得更为明显,尤其在样本量减少的时候。

B. 当项目数为 40 时 此时需要在三个组中分别比较三角形和方块折线。观察可知,两种方法参数估计的变化模式与项目数为 15 时类似。只是对于 α 的估计在中小样本时两种方法的差异略有增大,对于 γ 的估计,在中小样本时两种方法的差异略有减小。

3.3.3 比较项目数 J 对参数估计的影响 可以从图 1 直接观察。观察项目数对 DA-T Gibbs 的影响,只需考虑圆圈折线($J = 15$)与三角形折线($J = 40$)的比较;观察项目数对 M-H within Gibbs 的影响,只需考虑加号折线($J = 15$)与方块折线($J = 40$)的比较。图中分为了四组,每组对应于不同的样本量。

在各种样本量下,随着项目数的增加,DA-T Gibbs 方法的 α 参数的估计均变得更好, γ 参数的估计变化不大。同时,随着缺失比例的增加, α 参数的 RMSE 降低的幅度也增加,即 α 参数变好的幅度增加。

对于 M-H within Gibbs 方法,在样本量较大且缺失比例较小时,随着题目数的增加,参数估计几乎不变;其余情况则是随着题目数的增加,估计均变得更好,且随着样本量的减少或缺失比例的增加,改善的幅度也有所增加。

因此,在样本量减少或缺失比例增加时,增加题目数均可以改善两种方法的参数估计。

4 实证研究

4.1 研究设计

为了验证模拟研究的结果在实际数据中的稳定性,我们运用了实证研究——2006 年宜昌市中考数

学试题研究。该试题中一共有 15 道二分题目(10 道选择题,5 道填空题)。数据没有缺失,为比较两种 MCMC 方法对缺失数据参数估计的精确性,我们对这个真实的反应矩阵按上一小节模拟缺失的方法分别按 $p=0.05, 0.1, 0.25$ 三种缺失比例模拟缺失,得到含有缺失数据的反应矩阵,再运用两种 MCMC 方法。

同时,我们对完整反应矩阵(无缺失)用 PARSCALE 软件估计出项目参数,把该估计作为参数的“真实值”,再把两种 MCMC 方法对各种缺失条件下的缺失数据估计的参数与该“真实值”进行比较。仍然运用模拟研究中所用指标 RMSE,只是此时指

标中的真实值实际为无缺失时 PARSCALE 软件的估计结果。

4.2 结果

分别运用两种 MCMC 方法进行参数估计,得到的 RMSE 结果见表 3。

从表中结果可以看出,当缺失比例从 0.05 增加到 0.1 时,两种方法参数估计的 RMSE 几乎没有变化,而缺失比例从 0.1 增加到 0.25 时,两种方法的参数估计都变坏。不管缺失比例是多少,均有 DA-T Gibbs 方法优于 MH within Gibbs 方法。这与模拟结果是基本一致的。

表 3 实证研究中两种方法的 RMSE

项目数 J	被试人数 N	缺失比例 p	RMSE			
			斜率参数 α		截距参数 γ	
			MH within Gibbs	DA-T Gibbs	MH within Gibbs	DA-T Gibbs
15	810	0.05	0.24	0.15	0.21	0.13
		0.10	0.24	0.14	0.19	0.11
		0.25	0.28	0.23	0.23	0.18

5 结论与讨论

MCMC 方法是处理 2PL 模型缺失数据的一种可行有效的方法。本文在不同被试人数,测验长度和缺失比例下比较了两种 MCMC 缺失数据处理方法的精确度,为今后不同情况下的实际应用提供了方法选择的指导。18 种组合情况相对更好的方法可以直接由图 1 读出(RMSE 更小的方法更好),从而可以根据该图选择参数估计方法。

由本文的研究可知两种方法的项目参数估计均受被试人数影响很大,受缺失比例影响相对更小(在一定的缺失比例范围内)。M-H within Gibbs 受样本量的影响大于 DA-T Gibbs,受缺失比例的影响略大于 DA-T Gibbs。在大样本小缺失的时候,M-H within Gibbs 略优于 DA-T Gibbs;随着样本量的减少或缺失比例的增加,DA-T Gibbs 会反过来优于 M-H within Gibbs,且随着样本量的减少或缺失比例的增加,优势变得更为明显,尤其在样本量减少的时候。因此在样本量较小或缺失比例较大时,更应选择的是 DA-T Gibbs 方法,尤其样本量较少时。同时,在样本量减少或缺失比例增加时,增加题目数均可以改善两种方法的参数估计,但这对样本量较大且缺失比例较小时的 M-H within Gibbs 是几乎没有作用的。

此外,由于 DA-T Gibbs 方法增加了未知变量且

抽样过程需要对该未知变量抽样,因此 DA-T Gibbs 方法的运行速度会比 M-H within Gibbs 更慢,这是这一方法的一个缺点。因此,在大样本小缺失时建议使用 M-H within Gibbs 方法,在小样本或缺失比例较大时建议使用 DA-T Gibbs 方法。

MCMC 方法用样本分布代替了真实后验分布,简化了计算,得到的估计也不错。但对后验分布进行抽样时,马尔科夫链要达到一定的长度才能收敛。而这个长度的选择还没有一个很合适的方法。本文采取的方法,是取一个足够长的链,即迭代抽样 8000 次,以保证链收敛。但这样做的缺点是需要花费的时间较长。

本文的模拟研究产生的模拟缺失数据是基于完全随机缺失机制(Little & Rubin, 2002/2004),进一步的比较还可以基于随机缺失机制(Little & Rubin, 2002/2004)。

参 考 文 献

- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17, 251-269.
- Huisman, M. (2000). Imputation of missing item responses: some simple techniques. *Quality and Quantity*, 34, 331-351.
- Jones, D. H., & Nediak, M. (2000). Item parameter calibration of LSAT items using MCMC approximation of Bayes posterior distributions. *Rutcor research report*, 7-2000.
- Karkee, T., & Finkelman, M. (2007, April). *Missing Data treatment Methods in Parameter Recovery for a Mixed-Format Test*.

- Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Little, R. J. A. , & Rubin, D. B. (2004). *Statistical analysis with missing data*. 2nd ed. (Sun, S. Z. , Trans.). New York: John Wiley & Sons. (Original work published 2002)
- [利特尔, 鲁宾. (2004). 缺失数据统计分析 (孙山泽 译). 北京: 中国统计出版社.]
- Ludlow, L. H. , & O'Leary M. (1999). Scoring omitted and not-reached items: practical data analysis implications. *Educational and Psychological Measurement*, 59, 615–630.
- Maris, G. , & Bechger, T. M. (2005). An introduction to the DA-T Gibbs sampler for the two-parameter logistic (2PL) model and beyond. *Psicol gica*, 26, 327–352.
- Muraki, E. , & Bock, R. D. (1993). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago: Scientific software Int.
- Patz, R. J. , & Junker, B. W. (1999a). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24, 146–178.
- Patz, R. J. , & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT; multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342–366.
- Qi, S. Q. , Dai, H. Q. , & Ding, S. L. (2002). *Principles of modern educational and psychological measurement (in Chinese)*. Beijing, China: Higher Education Press.
- [漆书青, 戴海琦, 丁树良. (2002). 现代教育与心理测量学原理 北京: 高等教育出版社.]
- Qi, S. Q. , Dai, H. Q. , & Ding, S. L. (2003). Comparison of ANOTE, MULTILOG and PARSCALE on the efficiency of parameter estimation (in Chinese). *Examinations Research*, 3.
- [漆书青, 戴海琦, 丁树良. (2003). IRT 分析程序 ANOTE 与 MULTILOG, PARSCALE 参数估计性能比较. 考试研究, 第三辑.]
- Ross, S. M. (2006). *Simulation (4th ed.)*. Beijing, China: Post & Telecom Press.
- Tsutakawa, R. K. (1985). 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* (pp. 180–188).

Comparison of Two MCMC Approaches to Missing Response Data in 2PL Model

ZENG Li¹, XIN Tao², ZHANG Shu-Mei¹

(¹ School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems, Ministry of Education, Beijing 100875, China) (² School of Psychology, Beijing Normal University, Beijing 100875, China)

Abstract

Missing response data is common in educational assessment surveys. MCMC (Markov chain Monte Carlo) is a method of simulating random samples from any theoretical multivariate distribution-in particular, from the multivariate posterior distribution that is the focus of Bayesian inference. So it can obtain Bayes parameter estimation using the simulated sample, such as the mean of the simulated sample of posterior distribution can be used to estimate the EAP (expected a posteriori) of parameters. This algorithm is easy to implement when the IRT model is complex and the most important is that it can deal with missing data easily. In the past studies involving MCMC algorithm in IRT, researchers just compared MCMC method with other algorithms, no one compared different MCMC approaches. But in fact, there are vary MCMC algorithms, and some of them have been implemented in IRT. So, if the different MCMC approaches in IRT perform same is needed to be explored.

This paper explored the relative performance of two different MCMC approaches: MH within Gibbs and DA-T Gibbs Sampler in the estimation of the two-parameter logistic (2PL) item parameters. Simulation studies and real data examples were used in the comparison. Within the simulation, the factor effects of sample size, test length and missing rate were investigated. We considered three different sample size (100, 500, 1000), two different test length (15, 40) and three different missing rate (0.05, 0.1, 0.25). So there are 18 combinatorial situations. In each situations, we generated 20 subject response matrix, used the two methods to estimate the item parameters and then used the index RMSE (root mean square error) to compare the two approaches.

The simulation study results showed that the two MCMC approaches were indeed different in item parameter estimation. The parameter estimations of the two methods were both affected by sample size significantly, while the effect of the missing rate was relatively small. When the sample size is large and the missing rate is small, the RMSE of parameter estimation in MH within Gibbs is relatively small, and as the decreasing of sample size or the increasing of missing rate, DA-T Gibbs Sampler became better than MH within Gibbs. The results of real data example are consistent with the simulation results.

Key words IRT; missing data; MCMC