

# 基于小样本容量的 IRT 参数估计方法比较研究<sup>\*</sup>

朱隆尹<sup>1,2</sup>, 丁树良<sup>1</sup>, 涂冬波<sup>3</sup>, 卢震辉<sup>2</sup>

(1. 江西师范大学 计算机信息工程学院, 南昌 330022; 2. 赣南师范学院 数学与计算机科学学院, 赣州 341000;

3. 江西师范大学 心理学院, 南昌 330022)

**摘 要:**目前 IRT 参数估计方法的研究大多数都是基于大样本容量,而基于小样本容量的 IRT 参数估计方法的研究较少。通过对 IRT 参数估计方法的比较分析给出了基于小样本容量的 IRT 参数估计方法。实验数据表明,恰当的估计方法(如改进初值的 BP 神经网络)能大大提高小样本的参数估计精度。

**关键词:**CMLE;MMLE/EM;EAPE;MCMC;BP;蒙特卡洛模拟

**中图分类号:**B841.2

**文献标识码:**A

**文章编号:**1003-5184(2009)05-0072-05

## 1 引言

项目反应理论(Item Response Theory,简称 IRT)是现代教育与心理测量理论之一,与经典测量理论相比,IRT 具有许多优点<sup>[1-6]</sup>,但它同时也具有参数估计相对困难的特点。目前已有许多学者对 IRT 参数估计方法进行了研究,但绝大多数都是在大本容量的基础上,而专门针对小样本容量的参数估计方法研究则很少。从数理统计理论来讲,样本容量越大,IRT 参数估计结果就越精确。但在实际测验中,小样本容量的情况也经常出现。比如要对某些特殊部门或特殊专业人员进行考试,或者等值设计中因为考试成本太高,不得不采用比较少的考生。当然,其中一个解决办法就是把若干个小样本结合起来汇成大样本数据。然而,在实际测验中有时这种方法却难以实现,比如专业测验中,常常因为专业学生人数很少,大样本容量成了几乎是可望而不可及的奢求。若将几年的测验数据汇聚在一起,又不能保证被试群体能力分布稳定。如何对这些小样本容量的 IRT 参数进行估计就成了一个非常现实和重要的需求。目前已有一些学者在研究 IRT 参数估计技术的同时也对小样本容量进行了一些探索,但一般都是顺带提及,目前专门系统针对小样本容量的 IRT 参数估计研究的文献不多。

不同的学科和文献对小样本有不同的界定,文献<sup>[2]</sup>和文献<sup>[16]</sup>指出 2PLM 下用 BILOG 估计的项目数不低于 30 个,被试数不低于 500。为此文章定义小样本的涵义是项目数低于 30 个或者被试数量低于 500 的样本。

IRT 将被试特质水平与被试在项目上的行为关联起来并且将其参数化、模型化<sup>[1-7]</sup>。IRT 模型有很多,对于 0~1 评分情形,一般采用逻辑斯蒂克(Logistic)模型,Logistic 模型有单参数、双参数和三参数模型之分。而双参数 Logistic 模型(2PLM)既相对简单又很有代表性,为简单起见,这里只对 2PLM 的项目参数和能力参数估计进行讨论。2PLM 模型如下:

$$P(\theta; a, b) = 1 / (1 + \exp(-Da(\theta - b)))$$

P 表示能力为  $\theta$  的被试答对区分度为  $a$  和难度为  $b$  的项目的概率。

IRT 模型有一组假定<sup>[1-5]</sup>:1) 每个被试的作答模式是相互独立的;2) 同一被试对各个项目的作答是相互独立的。设有  $N$  个被试参加长度为  $m$  的测验,第  $i$  个被试能力记为  $\theta_i, i=1, \dots, N$ ;第  $j$  个项目的难度、区分度分别记为  $b_j, a_j, j=1, \dots, m$ ;被试  $i$  对第  $j$  个项目的作答反应记为  $u_{ij}$ (答对为 1,答错为 0)。

## 2 参数估计方法分析

2.1 条件极大似然估计(Condition Maximum Likelihood Estimation,简称 CMLE)和联合极大似然估计(Joint Maximum Likelihood Estimation,简称 JMLE)

因对不同的  $\theta$  和不同的  $j, u_{ij}$  的分布不尽相同,所以无法采用矩估计方法<sup>[3]</sup>。一般采用极大似然估计方法,当已知项目参数(或能力参数)而估计能力参数(或项目参数)时就是条件极大似然估计。根据要估计的参数 CMLE 可分为能力 CMLE 和项目参数 CMLE,其具体做法请参见文献<sup>[1]~[5]</sup>。

CMLE 方法的前提是必须先已知能力参数或项

<sup>\*</sup> 基金项目:国家自然科学基金(30860084),江西省教育厅科技项目基金(GJ08154),卫生部项目(KY200704),教电馆研(063120140, 063120144),江西省教学改革项目(JXJG-08-60-2)。

目参数(事实上这要求往往难以满足),另外它不能处理全部答对(错)的特殊情况。

因为CMLE方法是在已知能力参数或项目参数情况下使用,而实际测验时往往是两种参数都未知,这时就提出了JMLE<sup>[1-5]</sup>。采用JMLE时,未知能力参数的个数随着被试人数的增加而增加,这种未知参数称为伴随参数<sup>[1-5]</sup>,伴随参数的出现往往导致项目参数的MLE不一定是相合估计,这事实上也是JMLE的一个固有缺陷<sup>[1-5]</sup>。

2.2 和边际极大似然估计 EM 算法(Marginal Maximum Likelihood Estimation and an EM algorithm,简称MMLE/EM)

后来有学者给出能力的先验分布对能力参数积分,采用边际化手段来消除能力参数,即给出项目参数的边际极大似然估计(MMLE),这样就消除了JMLE的固有缺陷。

MMLE因迭代算法的局限性使得其只适应m较小的场合<sup>[1-5]</sup>。Bock和Aitkin把EM算法和MMLE结合在一起形成了MMLE/EM算法<sup>[1-5]</sup>,该算法分为两步:E步(求期望)和M步(求极大)。

该算法的优点是消除了能力参数的影响,缺点是要先给定能力先验分布和不能处理被试全部答对(错)的特殊情况。

2.3 能力贝叶斯期望后验估计(Bayesian Expected a Posterior Estimation,简称EAPE)<sup>[1-5]</sup>

EAPE的数值积分形式为:

$$E(u, \theta) = \frac{\int \prod_{k=1}^q X_k L(X_k) A(X_k) L(X_k) A(X_k) dX_k}{\int \prod_{k=1}^q L(X_k) A(X_k) dX_k} \quad (1)$$

该算法的优点是不需要迭代和能处理被试全部答对(错)的特殊情况,缺点是若能力先验分布与真实分布相差较大,而先验分布的方差又较小时,估计出来的能力会向先验分布的能力均值靠近,就会发生较大偏差。

2.4 马尔可夫链蒙特卡洛估计方法(Markov Chain Monte Carlo,简称MCMC)

MCMC方法是一种动态的计算机模拟技术,它是根据任一多元理论分布,特别是根据以贝叶斯推断为中心的多元后验分布来模拟随机样本的一种方法<sup>[9]</sup>。文献[8]~[12]给出了MCMC算法具体实现方法和步骤。

前面3种参数估计方法都对初值比较敏感,初值选取不当可能会导致迭代失败。但MCMC算法相对前面方法而言对初值敏感大大减小,但缺点是估计时间较长和要先给定参数建议分布。

对IRT中未知参数还有神经网络估计方法。

2.5 BP神经网络(Proposed distribution)估计法(简称BP)

BP前向神经网络是一个将W-H学习规则一般化,对非线性可微分函数进行权值训练的多层网络<sup>[15]</sup>。

IRT模型中项目参数区分度a的BP估计具体算法如下<sup>[13,14]</sup>:

1)模拟生成M个项目参数,并根据这些模拟的项目参数值和已知的N个能力参数值模拟产生30个得分矩阵<sup>[1-5]</sup>。

2)对每个得分矩阵用题分与总分的相关系数将得分阵降成1行m列的向量。这样30个得分矩阵就降维产生了30个向量,把这30个向量分别作为30个神经网络的输入,训练这30个神经网络。

3)对准备要估计的实际得分阵用题分与总分的相关系数把得分阵降成1行m列的向量作为测试网络的输入向量。

4)测试原来训练好的30个网络,以模拟项目参数为网络目标值,每次网络输出进行相加,循环30次后,所得的平均值就是所要的项目参数区分度估计值。

项目参数难度b的BP估计算法只要把上面算法中第2)步中题分与总分相关系数改为项目通过率即可<sup>[13,14]</sup>,由于这种将降维方法使用到BP神经网络的方法首次见于谭云兰等<sup>[13,14]</sup>的文章,故称这种BP方法为谭氏BP法。

上面的BP估计法至少有二个要点:一是已知能力参数;二是对输入向量采用降维技术。对第一点,如果真的是要必须先已知能力参数,则大大缩小BP的应用范围,因为测验实践中能力参数的获得往往是一个较困难的事,文献[14]指出可采用文献[1]~[5]中求能力初值的方法给出能力估计初值向量,这样就大大拓宽了BP的应用范围。

在小样本容量的前提下,对BP估计法分两方面进行改进。事实上用文献[1]~[5]的方法求得的能力初值非常粗糙,注意到用MMLE/EM算法和EAPE估计后得到的项目参数和能力参数远比[1]~[5]介绍的求初值方法得到的结果要精确得多,故可考虑对BP估计进行修正。

修正的项目参数区分度a的BP估计方法为把谭氏BP中的第2)、3)步的降维方法改为用MMLE/EM估计的项目区分度参数值,这里称为新BP法。

对难度BP估计用两个方法来修正:

第一个修正方法(称为新BP1法)是对谭氏BP

的第 1) 步能力参数由文献 [1] ~ [5] 的初值方法改为先用 MMLE/ EM 估计项目参数再用 EAPE 估计能力参数。

第二个修正方法 (称为新 BP2 法) 在新 BP1 法的基础上把项目通过率的降维技术改为用 MMLE/ EM 估计项目难度技术。

BP 估计法的优点是可处理被试全部答对 (错) 的特殊情况, 缺点是估计时间较长和先获知项目参数的先验分布。

3 实验与结论

3.1 算法评价标准

前面已讲述适合小样本容量的几种估计方法。考虑到参数估计算法主要关心的还是估计精度, 为此引入下面二个评价指标:

估计绝对偏差  $ABS = \frac{1}{m} \sum_{j=1}^m | \hat{\theta}_j - \theta_j | \dots\dots (2)$

估计标准误差  $MSE = \frac{1}{m} \sum_{j=1}^m ( \hat{\theta}_j - \theta_j )^2 \dots\dots (3)$

表 1 区分度 a 的 BP 估计方法比较结果

实验情景	方法	Lna ~ N(0,1) ,b ~ N(0,1)		Lna ~ N(0,1) ,b ~ U(- 3,3)	
		ABS	MSE	ABS	MSE
50 人 20 题	谭氏 BP	0.313262	0.181252	0.384225	0.237109
	新 BP	0.311602	0.166795	0.357625	0.230811
300 人 20 题	谭氏 BP	0.246468	0.116558	0.445701	0.293938
	新 BP	0.180528	0.058767	0.314826	0.20191
50 人 40 题	谭氏 BP	0.302192	0.159113	0.359165	0.198713
	新 BP	0.280316	0.137011	0.31785	0.172773
300 人 40 题	谭氏 BP	0.208087	0.087233	0.399913	0.231752
	新 BP	0.173159	0.058732	0.292279	0.158448
600 人 20 题	谭氏 BP	0.229495	0.118093	0.423144	0.248671
	新 BP	0.169793	0.059634	0.248614	0.132251

2) 实验二: 项目难度 b 的最佳 BP 估计法探索  
实验是用来比较谭氏 BP 和两种修正方法的优劣, 结果见表 2。

表 2 难度 b 的 BP 估计方法比较结果

实验情景	方法	Lna ~ N(0,1) ,b ~ N(0,1)		a ~ U(0.2,2.5) ,b ~ U(- 3,3)	
		ABS	MSE	ABS	MSE
50 人 20 题	谭氏 BP	0.337978	0.180743	0.635619	0.61976
	新 BP1	0.300642	0.145106	0.400945	0.268908
	新 BP2	0.320155	0.206155	0.493162	0.491461
300 人 20 题	谭氏 BP	0.270819	0.129402	0.584537	0.516948
	新 BP1	0.223022	0.096248	0.332216	0.190641
	新 BP2	0.19203	0.068051	0.396881	0.293157
50 人 40 题	谭氏 BP	0.303911	0.162422	0.516761	0.419892
	新 BP1	0.299685	0.155246	0.372574	0.239766
	新 BP2	0.319453	0.197968	0.42541	0.386696
300 人 40 题	谭氏 BP	0.241118	0.100166	0.572344	0.502287
	新 BP1	0.22694	0.090623	0.329908	0.198042
	新 BP2	0.18183	0.061075	0.370839	0.225661
600 人 20 题	谭氏 BP	0.260483	0.116867	0.586213	0.516861
	新 BP1	0.215216	0.080823	0.311335	0.185613
	新 BP2	0.183016	0.052435	0.401505	0.276556

从表中结果可以看出修正 BP 方法有明显的优势,在  $a \sim U(0.2, 2.5)$ ,  $b \sim U(-3, 3)$  时新 BP1 比新 BP2 要好,其他情形下人少 (50 人) 时新 BP1 比新 BP2 好,人多时新 BP2 比新 BP1 要好。

3) 实验三:不同估计方法的比较  
实验三用来比较不同的参数估计方法,其中 EAPE 能力估计法中的项目参数取自 MMLE/EM 算法中的项目参数,BP 算法中的结果是取前两个实验的最好实验结果,同时还用 BILOG 程序进行了估计,实验结果如表 3、4。

从上述表中可看出 MCMC 估计法具有优势,特别是人少 (50 人) 时尤其明显。但在 600 人 20 题时难度估计 BILOG 略比 MCMC 好些。

表 3 小样本不同参数估计方法比较结果 ( $Lna \sim N(0, 1)$ ,  $b \sim N(0, 1)$ )

评价指标		方法	50 人 20 题	300 人 20 题	50 人 40 题	300 人 40 题	600 人 20 题
区分度	ABS	MMLE/EM	0.3512	0.1626	0.3134	0.1627	0.1411
		BP	0.3116	0.1805	0.2803	0.1732	0.1698
		MCMC	0.2570	0.1503	0.2454	0.1501	0.1233
		BILOG	0.2787	0.1287	0.2610	0.1279	0.1117
	MSE	MMLE/EM	0.2441	0.0582	0.1960	0.0631	0.0545
		BP	0.1668	0.0588	0.1370	0.0587	0.0596
		MCMC	0.1186	0.0420	0.1134	0.0390	0.0297
		BILOG	0.1314	0.0327	0.1156	0.0331	0.0278
难度	ABS	MMLE/EM	0.4141	0.1960	0.4118	0.1977	0.1584
		BP	0.3006	0.1920	0.2997	0.1818	0.1830
		MCMC	0.2953	0.1410	0.2958	0.1444	0.1148
		BILOG	0.2883	0.1454	0.3041	0.1590	0.1012
	MSE	MMLE/EM	0.3815	0.0821	0.4025	0.0797	0.0477
		BP	0.1451	0.0681	0.1552	0.0611	0.0524
		MCMC	0.1373	0.0421	0.1447	0.0407	0.0239
		BILOG	0.1364	0.0425	0.1586	0.0454	0.0201
能力	ABS	EAPE	0.3278	0.2999	0.2598	0.2359	0.2996
		MCMC	0.3065	0.2868	0.2422	0.2213	0.2860
		BILOG	0.3209	0.2946	0.2688	0.2227	0.2893
	MSE	EAPE	0.1714	0.1450	0.1087	0.0901	0.1421
		MCMC	0.1504	0.1366	0.0947	0.0820	0.1323
		BILOG	0.1639	0.1405	0.1152	0.0811	0.1331

表 4 小样本不同参数估计方法比较结果 ( $a \sim U(0.2, 2.5)$ ,  $b \sim N(0, 1)$ )

评价指标		方法	50 人 20 题	300 人 20 题	50 人 40 题	300 人 40 题	600 人 20 题
区分度	ABS	MMLE/EM	0.4604	0.2651	0.4441	0.2996	0.201976
		BP	0.3752	0.2818	0.3814	0.2706	0.315598
		MCMC	0.3146	0.1580	0.3000	0.1532	0.141186
		BILOG	0.3557	0.2138	0.3438	0.2639	0.174175
	MSE	MMLE/EM	0.4011	0.1578	0.3773	0.2016	0.098503
		BP	0.2273	0.1372	0.2295	0.1412	0.15389
		MCMC	0.1704	0.0439	0.1521	0.0431	0.037838
		BILOG	0.2205	0.0881	0.2101	0.1203	0.062408
难度	ABS	MMLE/EM	0.4173	0.2020	0.4212	0.2330	0.170586
		BP	0.2674	0.2086	0.2781	0.1861	0.233158
		MCMC	0.2531	0.1468	0.2607	0.1389	0.13824
		BILOG	0.2826	0.1618	0.2880	0.1948	0.130599
	MSE	MMLE/EM	0.4016	0.0931	0.3971	0.1187	0.055943
		BP	0.1211	0.0730	0.1294	0.0627	0.083573
		MCMC	0.1342	0.0383	0.1438	0.0339	0.032233
		BILOG	0.1451	0.0563	0.1411	0.0670	0.034733
能力	ABS	EAPE	0.3304	0.2845	0.2855	0.2363	0.282474
		MCMC	0.2812	0.2597	0.2110	0.1989	0.269502
		BILOG	0.3219	0.2762	0.2509	0.2353	0.274128
	MSE	EAPE	0.1723	0.1291	0.1324	0.0917	0.127274
		MCMC	0.1305	0.1111	0.0742	0.0649	0.117957
		BILOG	0.1630	0.1225	0.0994	0.0908	0.120195

#### 4 进一步要研究的问题

这里只研究了 2PLM 模型下的小样本估计方法的分析比较,这些结论能否推广到其他模型中是今后进一步要研究的问题。

这里通过适合小样本的参数估计的比较分析,得出了 2PLM 模型下 MCMC 估计法比较适合小样本容量的参数估计,对拓宽 IRT 模型的适用范围有一定的价值。

#### 参考文献

- 1 漆书青,戴海崎,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.
- 2 Baker FB, Kim S H. Item Response Theory: Parameter Estimate Technique. 2<sup>nd</sup>. ed. Marcel Dekker, 2004.
- 3 Hambleton R K, Swaminathan H. Item Response Theory: Principles and Applications. Boston: Kluwer - Nijhoff Publishing, 1985.
- 4 Lord F M. Applications of Item Response Theory to practical testing problems. Hillsdale N J: Lawrence Erlbaum Associates Inc, 1980.
- 5 漆书青. 现代测量理论在考试中的应用. 武汉: 华中师范大学出版社, 2003.
- 6 罗芬. 项目反应理论中 2PLM 参数估计新方法. 硕士学位论文. 南昌: 江西师范大学, 2003.
- 7 罗芬, 丁树良, 胡小松, 等. 基于 IRT 若干参数估计方式的比较. 江西师范大学学报(自然科学版), 2003, (1): 56 - 60.
- 8 涂冬波, 漆书青, 蔡艳, 等. IRT 模型参数估计的新方法 - MCMC. 心理科学, 2008, (1).
- 9 王权. “马尔可夫链蒙特卡洛”(MCMC) 方法在估计 IRT 模型参数中的应用. 考试研究, 2006, 2(4): 45 - 63.
- 10 Richard J, Patz, Brian W, et al. A straightforward approach to Markov Chain Monte Carlo Methods for Item Response Models. Journal of Educational and Behavioral Statistics, 1999, 24(2): 146 - 178.
- 11 Richard J, Patz, Brian W, et al. Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics, 1999, 24(4): 342 - 366.
- 12 Richard J, Patz, Brian W, Junker, et al. The Hierarchical Rater Model for Rated Test items and its application to Large - scale Educational Assessment Data. Journal of Educational and Behavioral Statistics, 2002, 27(4): 341 - 384.
- 13 谭云兰, 丁树良, 辛锐铭, 等. 基于 IRT 模型参数的 BP 神经网络估计. 计算机工程与应用, 2004, 17.
- 14 谭云兰, 丁树良, 辛锐铭. 基于 IRT 模型的 BP 神经网络降维参数估计及其应用. 江西师范大学学报(自然科学版), 2004, 6.
- 15 从爽. 面向 MATLAB 工具箱的神经网络理论与应用. 第 2 版. 北京: 中国科学技术大学出版社, 2003.
- 16 张权. 语言测试中的项目分析与等值技术: 研究与应用. 北京: 高等教育出版社, 2004.

## Comparison among Parameter Estimation Methods Based on Small Sample under Item Response Theory

Zhu Longyin<sup>1,2</sup>, Ding Shuliang<sup>1</sup>, Tu Dongbo<sup>3</sup>, Lu Zhenhui<sup>2</sup>

(1. Jiangxi Normal University, Nanchang 330022; 2. Gannan Normal University, Ganzhou 341000;

3. School of Psychology, Jiangxi Normal University, Nanchang 330022)

**Abstract:** At present, the study of IRT parameter estimation method is mostly grounded on the large sample, while the study of IRT parameter estimation method based on the small sample is less. In this paper IRT parameter estimation methods based on the small sample are proposed and comparison among these methods is made. The results of the experiment show that the accuracy of the parameter estimation on small scale sample can be greatly improved with the proper estimation method, such as the BP neural network with improving initial value.

**Key words:** CMIE; MML/EM; EAPE; MCMC; BP; Monte Carlo simulation



知网查重限时 **7折** 最高可优惠 **120元**

本科定稿，硕博定稿，查重结果与学校一致

立即检测

免费论文查重: <http://www.paperyy.com>

3亿免费文献下载: <http://www.ixueshu.com>

超值论文自动降重: [http://www.paperyy.com/reduce\\_repetition](http://www.paperyy.com/reduce_repetition)

PPT免费模版下载: <http://ppt.ixueshu.com>

---