

基于 MCMC 三种抽样方法的概化理论 方差分量估计准确性比较*

黎光明

(华南师范大学心理学院、心理应用研究中心, 广州 510631)

摘要 概化理论是关于行为测量可靠性的现代心理测验理论, 已被广泛应用于心理测量实践中。基于 $p \times i$ 和 $p \times i \times h$ 设计, 采用模拟研究和实证研究, 使用 MCMC 三种抽样方法(M-H 算法、Gibbs 抽样和 HMC 算法)对概化理论方差分量进行估计, 并比较其差异。结果表明: (1)少量数据缺失对 MCMC 三种抽样方法几乎无影响, MCMC 对于缺失数据估计具有较强的稳健性; (2)整体上, 有信息先验和经验性先验较无信息先验估计准确性更高; (3)仅 HMC 算法表现出“跨设计性”, 在两种设计之下的估计准确性都较高, 但 M-H 算法和 Gibbs 抽样却受限于 $\sigma^2(h)$, 估计准确性都不高; (4)若 M-H 算法和 Gibbs 抽样需要克服 $\sigma^2(h)$ 估计准确性不高的局限性, 则需规定 $h \geq 4$; 未来研究者在使用 MCMC 方法时, HMC 算法更值得推荐。

关键词 概化理论, 方差分量估计, MCMC 方法, M-H 算法, Gibbs 抽样, HMC 算法

分类号 B841

1 引言

概化理论(Generalizability Theory, GT)是在经典测验理论基础上发展起来的, 是关于行为测量可靠性的理论(黎光明, 秦越, 2022)。GT 能够区分出测量过程中的多种误差来源, 并估计不同来源误差对测量目标的影响, 已被广泛应用于心理测量的实践中, 如划界分数设定、学业水平考试、表现性评价、人才选拔测评等诸多方面(赵轩 等, 2018; 郑雪琴 等, 2025; Jackson et al., 2016; Li & Wu, 2023)。

概化理论分析包括概化研究(Generalizability study, G 研究)和决策研究(Decision study, D 研究)两个过程(杨志明, 张雷, 2003; Liao, 2023)。G 研究的主要任务是尽可能多地挖掘测量过程中的各种误差来源, 并估计出不同误差来源的方差分量(Variance Components, VC); D 研究的主要任务是在概化研究的基础上, 根据实际中具体的决策需要,

通过调整测量过程中的各种关系来控制测量误差。由此可见, 方差分量的估计是 GT 分析的关键步骤(黎光明 等, 2022)。

概化研究中方差分量的估计受到多种因素的影响, 包括估计方法、数据分布、研究设计、样本容量等(黎光明, 2016; Brennan et al., 2022; Li, 2023)。Traditional、Bootstrap、Jackknife、MINQUE、REML (Restricted Maximum Likelihood Estimation) 和 MCMC (Markov chain Monte Carlo)等方法都可以用来估计方差分量(Brennan, 2001; Jackson et al., 2016; Lopilato et al., 2015)。前人研究表明, 将 MCMC 方法用于 GT 中估计方差分量具有以下优势(Chen et al., 2022; Enders et al., 2013; Jiang & Skorupski, 2018; Lopilato et al., 2015): 一是能提高方差分量估计的准确性, 并估计出 D 研究中概化系数和可靠性指数的置信区间, 还可以为多元概化理论提供协方差分量的标准误和置信区间估计值; 二

收稿日期: 2025-06-16

* 广东省自然科学基金面上项目(2021A1515012516)、广东省哲学社会科学规划学科共建项目(GD24XXL03)和华南师范大学冲补强心理学高峰学科项目。

通信作者: 黎光明, E-mail: Lgm2004100@m.scnu.edu.cn

是与其他估计方法相比, MCMC 方法不会估计出负的方差分量; 三是 MCMC 方法具有灵活性, 理论上能更好地处理缺失数据和小样本数据。

但是, MCMC 方法存在多种抽样方法, 需要进行比较。Metropolis-Hastings (M-H) 算法是 MCMC 方法中的核心抽样法(黄长全, 2017)。Gibbs 抽样是接受概率为 1 的 M-H 算法, 其简单易懂(刘金山, 夏强, 2016)。M-H 算法和 Gibbs 抽样都属于局部更新算法, 具有不可避免的随机游走行为; 当对复杂模型参数估计时, 这两种算法都可能存在计算量过大、耗费时间过长等问题, 尤其当模型参数之间存在较强的相关关系时, 收敛效果不太理想(Hoffman & Gelman, 2014)。后来, 有学者提出了一种新的抽样方法——HMC (Hamiltonian Monte Carlo) 算法, 能够有效避免抽样过程中的随机游走行为, 提高收敛的有效性(Neal, 2010)。有研究显示, 与 M-H 算法和 Gibbs 抽样相比, HMC 算法具有一些优势(Nugroho & Morimoto, 2015)。HMC 算法除了是一种全局更新算法, 能够有效避免随机游走行为之外, 其接受概率较高, 能够迅速收敛至目标后验分布, 还能处理高维积分的复杂模型。Lopilato 等(2015)最早使用 HMC 算法来估计 GT 中的方差分量, 并采用双侧面交叉设计比较了 REML 方法和 MCMC 方法的估计准确性, 但没有对 HMC 算法和其他的 MCMC 抽样方法的估计表现进行比较。目前, 对于 HMC 算法的研究主要集中于统计学方面和一些具体的应用研究(Bourabee & Sanzserna, 2017; Chaudhuri et al., 2017; Monnahan et al., 2017; Zhang et al., 2017)。但是, 其存在两点不足: 一是鲜有学者将 HMC 算法迁移至复杂的带有缺失数据的 GT 分析中; 二是 MCMC 三种抽样方法应用于 GT 分析中的性能如何? 迄今为止尚未有学者进行过探讨。

单侧面交叉设计指的是只有一个测量侧面, 测量目标和测量侧面之间存在交叉关系的测验设计, 即测量目标的每一个水平都要和测量侧面的每一个水平结合的测验设计。如果测量目标是 p , 测量侧面是 i , 那么单侧面交叉设计可用 $p \times i$ 表示。双侧面交叉设计中包含两个测量侧面, 并且每个测量侧面的每个水平都要和测量目标的每个水平一一结合。例如, 在九年级学生数学测试中, 每个学生都需要完成 5 个解答题, 并且每个学生的 5 个题目都被 3 位数学教师评价, 即每位教师都对每个学生在每一个解答题上进行了评分。如果测量目标是 p (学生), 两个测量侧面分别是 i (题目)和 h (教师), 那么

双侧面交叉设计可用 $p \times i \times h$ 表示。

本研究旨在将贝叶斯理论框架下的 MCMC 方法应用于 GT 分析中, 并比较 M-H 算法、Gibbs 抽样和 HMC 算法三种抽样方法, 在不同数据缺失比例下方差分量估计的性能差异。具体而言, 本研究基于 GT 中常见的两种概化设计($p \times i$ 设计和 $p \times i \times h$ 设计), 通过模拟研究和实证研究来探讨 MCMC 方法的三种抽样方法估计方差分量的准确性, 并深度探究这三种抽样方法估计方差分量的具体表现和使用范围, 为相应的实践提供理论指导。

2 单侧面交叉设计($p \times i$)模拟研究

在单侧面交叉设计 $p \times i$ 中, 分别使用 M-H 算法、Gibbs 抽样和 HMC 算法估计每种条件下的方差分量, 探讨三种抽样方法估计方差分量的表现及其影响因素。

2.1 研究设计

参考以往关于单侧面交叉设计 $p \times i$ 的设定(Mao et al., 2005), 将方差分量 $\sigma^2(p)$ 、 $\sigma^2(i)$ 和 $\sigma^2(pi)$ 的真值分别设为 4、16、64, 模拟正态分布的数据, 样本量固定为: p 的水平数是 100, i 的水平数是 20。自变量设置如下:

(1) MCMC 的三种抽样方法, 即 M-H 算法、Gibbs 抽样和 HMC 算法。

(2) 三种先验信息, 即有信息先验 MCMC (MCMC with informative priors, MCMC inf)、无信息先验 MCMC (MCMC with non-informative priors, MCMC non)和经验性先验 MCMC (MCMC with empirical priors, MCMC emp)。

根据贝叶斯统计学中共轭先验分布的性质(黄长全, 2017), 将先验信息的目标分布设为逆伽马分布(IGamma)。参考前人研究对先验信息的设定(Lopilato et al., 2015; Mao et al., 2005)。表 1 呈现了三种先验信息的具体设定情况。

(3) 三种完全随机缺失比例, 即 0%、5%和 10%。在心理统计与测量的实践过程中常会遇到数据缺失的情况, 根据前人研究可知, MCMC 方法能更好地处理缺失数据(Enders et al., 2013)。因此, 参考 Lopilato 等(2015)在其 GT 中对缺失比例的设置情况, 设定了三种完全随机缺失比例 0%、5%和 10%。

综上所述, 模拟的条件总数为: 3 种抽样方法(M-H、Gibbs、HMC) $\times$ 3 种先验信息(MCMC inf、MCMC non、MCMC emp) $\times$ 3 种缺失比例(0%、5%、10%) = 27。每种条件下都模拟 1000 次, 也就

表 1 单侧面交叉设计 $p \times i$ 的三种先验信息

先验信息	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
MCMC inf	IGamma(2, 4)	IGamma(2, 16)	IGamma(2, 64)
MCMC non	IGamma(0.001, 0.001)	IGamma(0.001, 0.001)	IGamma(0.001, 0.001)
MCMC emp	IGamma(2, 5.95)	IGamma(2, 24.07)	IGamma(2, 95.97)

是生成了 1000 批矩阵数据。方差分量的估计值取 1000 次估计结果的平均值。共有三个方差分量需要估计, 即 $\sigma^2(p)$ 、 $\sigma^2(i)$ 和 $\sigma^2(pi)$ 。

2.2 比较标准

比较标准选取估计偏差(Bias)和均方根误差(Root Mean Square Error, RMSE)(Medvedev & Siegert, 2022; Medvedeva et al., 2022), 其计算公式如下:

$$Bias = \frac{\sum(\hat{\theta} - \theta)}{n} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum(\hat{\theta} - \theta)^2}{n}} \quad (2)$$

在公式(1)和公式(2)中, θ 表示每次模拟统计量的估计值, 即方差分量的估计值; θ 表示统计量的真值, 即方差分量的真值; n 表示每种条件下的模拟次数。Bias 的绝对值越小, 估计越准确。RMSE 的值越小, 估计越准确。

2.3 数据模拟与估计

采用蒙特卡洛数据模拟技术, 在 R 软件中生成 1000 批正态分布的矩阵数据, 其依据的数学模型如下:

$$X_{pi} = \mu + \sigma_p z_p + \sigma_i z_i + \sigma_{pi} z_{pi} \quad (3)$$

在公式(3)中, 方差分量 $\sigma^2(p)$ 、 $\sigma^2(i)$ 和 $\sigma^2(pi)$ 的真值分别设为 4、16 和 64, μ 为 0, 需要模拟标准正态分布的 z_p 、 z_i 和 z_{pi} , 分别与 σ_p 、 σ_i 和 σ_{pi} 相乘, 根据公式再将 3 个乘积值相加得到模拟的数据(Mao et al., 2005)。在此基础上, 生成了样本量为 100×20 的完整数据矩阵。然后, 根据三种不同的缺失比例模拟出完全随机缺失下的缺失数据矩阵。

采用 M-H 算法、Gibbs 抽样和 HMC 算法, 分别对 $p \times i$ 设计数据进行估计, 得到相应的方差分量结果。其中, M-H 算法通过 R 软件调用 BRugs 包来实现; Gibbs 抽样通过 R 软件调用 R2jags 包与 JAGS 软件链接实现; HMC 算法通过 R 软件调用 Rstan 包来实现。MCMC 的每种算法都设置三条 Markov 链, 每条链迭代 10000 次, 在刚开始迭代时生成的随机数不稳定, 所以在研究中舍弃前 5000 次预热

(Burn-in)的迭代结果。

2.4 估计结果

对于 $p \times i$ 设计, 模型通过 R 软件进行抽样, 最终达到收敛。通过观察历史迭代图, 设定预热次数为 5000 就能达到平稳分布。因此, 从第 5001 开始至 10000 次, 每条 Markov 链的值都为估计值, 对三条 Markov 链求平均, 即为每个统计的估计值。三种缺失比例(缺失 0%/5%/10%)的 $p \times i$ 设计方差分量的估计结果如表 2 所示。

根据表 2 的信息, 可知 M-H 算法、Gibbs 抽样和 HMC 算法在三种先验信息条件下, 不同缺失比例的方差分量、Bias 和 RMSE 的估计结果。考虑到当缺失比例 0%和 5%时, 其结果与缺失比例 10%的结果基本相同, 在此不再列出, 仅以缺失比例 10%为例, 如图 1 所示。

从 Bias 的结果来看, 在有信息先验条件下(图 1 中的 a 图), M-H 算法、Gibbs 抽样和 HMC 算法的估计偏差范围为(-1, 0.4); 在经验性先验条件下(图 1 中的 c 图), 三者的估计偏差范围为(-0.2, 1); 而在无信息先验条件下(图 1 中的 b 图), 三者的估计偏差范围为(-0.5, 2.3)。在三种先验信息条件下, M-H 算法、Gibbs 抽样和 HMC 算法对 $\sigma^2(p)$ 的估计偏差都较为接近, 但对 $\sigma^2(i)$ 和 $\sigma^2(pi)$ 的估计偏差表现不一。在有信息先验条件下, M-H 算法对 $\sigma^2(i)$ 的估计偏差接近于 0, 而 Gibbs 抽样对 $\sigma^2(pi)$ 的估计偏差更接近于 0; 在经验性先验条件下, HMC 算法对 $\sigma^2(i)$ 的估计偏差接近于 0, 而 M-H 算法对 $\sigma^2(pi)$ 的估计偏差接近于 0; 在无信息先验条件下, 三种抽样方法对 $\sigma^2(pi)$ 的估计偏差较为接近, 而 HMC 算法对 $\sigma^2(i)$ 的估计偏差更小一点。

从 RMSE 的结果来看, M-H 算法、Gibbs 抽样和 HMC 算法在缺失比例为 10%、三种先验信息条件下, 估计 3 个方差分量的 RMSE 都比较接近。其中, $\sigma^2(p)$ 的 RMSE 值都接近于 1, $\sigma^2(pi)$ 的 RMSE 值都接近于 2, 而 $\sigma^2(i)$ 的 RMSE 值在有信息先验和经验性先验下都接近于 5, 在无信息先验条件下更接近于 6。

表 2 $p \times i$ 设计方差分量的估计结果(缺失 0%/5%/10%)

先验信息	抽样方法	估计值	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
MCMC inf	M-H	VC	3.797/3.815/3.743	15.664/15.727/15.868	64.059/64.121/64.167
		Bias	-0.203/-0.185/-0.257	-0.336/-0.273/-0.132	0.059/0.121/0.167
		RMSE	0.938/0.980/1.029	4.679/4.966/4.937	2.072/2.167/2.270
	Gibbs	VC	3.844/3.754/3.801	15.873/15.971/15.741	64.110/64.197/64.033
		Bias	-0.156/-0.246/-0.199	-0.127/-0.029/-0.259	0.110/0.197/0.033
		RMSE	1.025/0.989/1.053	4.929/5.082/4.725	2.106/2.123/2.264
	HMC	VC	3.625/3.694/3.683	15.156/15.112/15.199	64.124/64.095/64.173
		Bias	-0.375/-0.306/-0.317	-0.844/-0.888/-0.801	0.124/0.095/0.173
		RMSE	0.991/1.021/1.083	4.735/4.741/4.719	2.160/2.196/2.250
MCMC non	M-H	VC	4.015/4.014/4.003	17.664/17.683/17.835	63.966/64.156/64.175
		Bias	0.015/0.014/0.003	1.664/1.683/1.835	-0.034/0.156/0.175
		RMSE	1.068//1.068/1.086	6.023/6.219/6.321	2.047/2.162/2.189
	Gibbs	VC	4.065/4.048/3.957	17.787/17.678/18.031	64.063/64.138/64.114
		Bias	0.065/0.048/-0.043	1.787/1.678/2.031	0.063/0.138/0.114
		RMSE	1.086/1.079/1.138	6.261/6.086/6.501	2.037/2.175/2.272
	HMC	VC	3.897/3.813/3.763	16.390/16.909/16.758	64.111/64.074/64.190
		Bias	-0.103/-0.187/-0.237	0.390/0.909/0.758	0.111/0.074/0.190
		RMSE	1.132/1.157/1.215	5.585/5.914/5.754	2.170/2.157/2.214
MCMC emp	M-H	VC	3.980/3.931/3.963	16.639/16.760/16.814	64.048/64.155/64.022
		Bias	-0.020/-0.069/-0.037	0.639/0.760/0.814	0.048/0.155/0.022
		RMSE	0.964/0.908/0.943	4.747/4.851/4.921	2.016/2.162/2.185
	Gibbs	VC	3.990/3.900/3.898	16.449/16.836/16.611	64.096/64.056/64.195
		Bias	-0.010/-0.100/-0.102	0.449/0.836/0.611	0.096/0.056/0.195
		RMSE	0.951/0.943/0.965	4.612/4.917/4.822	2.064/2.177/2.236
	HMC	VC	3.836/3.818/3.853	16.053/15.814/16.032	64.111/64.111/64.182
		Bias	-0.164/-0.182/-0.147	0.053/-0.186/0.032	0.111/0.111/0.182
		RMSE	0.951/0.977/0.997	4.843/4.587/4.591	2.034/2.114/2.152

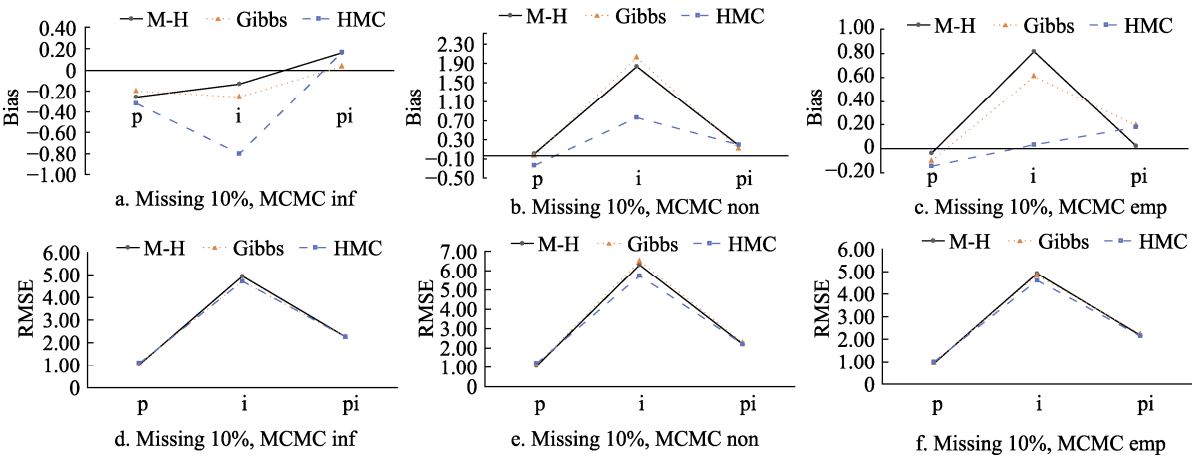


图 1 $p \times i$ 设计方差分量的 Bias 和 RMSE 图(缺失比例 10%)

总之，在单侧面交叉设计 $p \times i$ 中，M-H 算法、Gibbs 抽样和 HMC 算法的方差分量估计准确性较为一一致。不同缺失比例下，方差分量的估计表现差异很小。而三种先验信息条件下对方差分量的估计

准确性差异较小, 有信息先验和经验性先验条件下方差分量的估计准确性相对较高一些。

3 双侧面交叉设计($p \times i \times h$)模拟研究

在双侧面交叉设计 $p \times i \times h$ 中, 分别使用 M-H 算法、Gibbs 抽样和 HMC 算法估计每种条件下的方差分量, 探讨三种抽样方法估计方差分量的表现及其影响因素。

3.1 研究设计

参考以往关于双侧面交叉设计 $p \times i \times h$ 的设定(王幸君 等, 2016; Tong & Brennan, 2007), 将方差分量 $\sigma^2(p)$ 、 $\sigma^2(i)$ 、 $\sigma^2(h)$ 、 $\sigma^2(pi)$ 、 $\sigma^2(ph)$ 、 $\sigma^2(ih)$ 和 $\sigma^2(pih)$ 的真值分别设为 16、4、1、64、2、3 和 144, 模拟正态分布的数据, 样本量固定为: p 的水平数是 100, i 的水平数是 20, h 的水平数是 2。根据前人的研究(Lopilato et al., 2015), 有信息先验的设定值为方差分量的真值, 无信息先验的设定值为 $\text{IGamma}(0.001, 0.001)$, 经验性先验的设定值为 REML 方法对双侧面交叉设计中方差分量值估计值的 1.5 倍, 如表 3 所示。

与单侧面交叉设计 $p \times i$ 设计一样, 双侧面交叉设计 $p \times i \times h$ 缺失比例设置为 0%、5%和 10%, 共有 27 种模拟条件。每种条件下都模拟 1000 次, 这 1000 次估计结果的平均值即为方差分量的估计值, 共有 7 个方差分量需要估计, 即 $\sigma^2(p)$ 、 $\sigma^2(i)$ 、 $\sigma^2(h)$ 、 $\sigma^2(pi)$ 、 $\sigma^2(ph)$ 、 $\sigma^2(ih)$ 和 $\sigma^2(pih)$ 。

3.2 比较标准

同公式(1)和公式(2)。

3.3 数据模拟与估计

在 R 软件中生成 1000 批正态分布的矩阵数据, 其依据的数学模型如下:

$$X_{pi} = \mu + \sigma_p z_p + \sigma_i z_i + \sigma_h z_h + \sigma_{pi} z_{pi} + \sigma_{ph} z_{ph} + \sigma_{ih} z_{ih} + \sigma_{pih} z_{pih} \quad (4)$$

从公式(4)可知, 需要模拟正态分布的 7 个值, 即 z_p 、 z_i 、 z_h 、 z_{pi} 、 z_{ph} 、 z_{ih} 和 z_{pih} , 将这 7 个值分别与对应的方差分量 $\sigma^2(p)$ 、 $\sigma^2(i)$ 、 $\sigma^2(h)$ 、 $\sigma^2(pi)$ 、 $\sigma^2(ph)$ 、 $\sigma^2(ih)$ 、 $\sigma^2(pih)$ 的真值(16、4、1、64、2、3、144)相乘再求和, 就得到了模拟的随机数, 并生成样本量为 $100 \times 20 \times 2$ 的完整数据矩阵, 再根据三种不同的缺失比例模拟出完全随机缺失下的缺失数据矩阵。然后, 采用 M-H 算法、Gibbs 抽样和 HMC 算法分别对 $p \times i \times h$ 设计数据进行估计, 得到相应的方差分量估计值。其中, M-H 算法通过 R 软件调用 BRugs 包来实现; Gibbs 抽样通过 R 软件调用 R2OpenBUGS 与 BUGS 软件链接实现; HMC 算法通过 R 软件调用 Rstan 包来实现。每种抽样方法都设置三条 Markov 链, 每条链迭代 10000 次, 舍弃前 5000 次的迭代结果。

3.4 估计结果

3.4.1 不同缺失比例下的估计结果

与单侧面交叉设计 $p \times i$ 设计一样, $p \times i \times h$ 设计的概化模型通过 R 软件进行抽样, 最终达到收敛。通过观察历史迭代图, 设定预热次数为 5000 就能达到平稳分布, 因此从第 5001 开始至 10000 次, 每条 Markov 链的值都为估计值, 对三条 Markov 链求平均, 即为每个统计的估计值。 $p \times i \times h$ 设计方差分量的估计结果(缺失 0%、5%、10%)如表 4 至表 6 所示。

根据表 4、表 5 和表 6 的信息, 可知 M-H 算法、Gibbs 抽样和 HMC 算法在三种先验信息条件下, 不同缺失比例(0%、5%和 10%)的方差分量、Bias 和 RMSE 的估计结果。与单侧面交叉设计 $p \times i$ 设计一样, 仅以缺失比例 10%为例, 如图 2 所示。

从 Bias 的结果来看, 在有信息先验条件下(图 2 中的 a 图), M-H 算法、Gibbs 抽样和 HMC 算法的估计偏差范围为(-0.8, 1.2); 在经验性先验条件下(图 2 中的 c 图), 三者的估计偏差范围为(-0.8, 1.2);

表 3 双侧面交叉设计 $p \times i \times h$ 的三种先验信息

方差分量	MCMC inf	MCMC non	MCMC emp
$\sigma^2(p)$	$\text{IGamma}(2, 16)$	$\text{IGamma}(0.001, 0.001)$	$\text{IGamma}(2, 23.76)$
$\sigma^2(i)$	$\text{IGamma}(2, 4)$	$\text{IGamma}(0.001, 0.001)$	$\text{IGamma}(2, 6.03)$
$\sigma^2(h)$	$\text{IGamma}(2, 1)$	$\text{IGamma}(0.001, 0.001)$	$\text{IGamma}(2, 1.53)$
$\sigma^2(pi)$	$\text{IGamma}(2, 64)$	$\text{IGamma}(0.001, 0.001)$	$\text{IGamma}(2, 95.93)$
$\sigma^2(ph)$	$\text{IGamma}(2, 2)$	$\text{IGamma}(0.001, 0.001)$	$\text{IGamma}(2, 3.02)$
$\sigma^2(ih)$	$\text{IGamma}(2, 3)$	$\text{IGamma}(0.001, 0.001)$	$\text{IGamma}(2, 4.52)$
$\sigma^2(pih)$	$\text{IGamma}(2, 144)$	$\text{IGamma}(0.001, 0.001)$	$\text{IGamma}(2, 216.14)$

表 4 $p \times i \times h$ 设计方差分量的估计结果(缺失 0%)

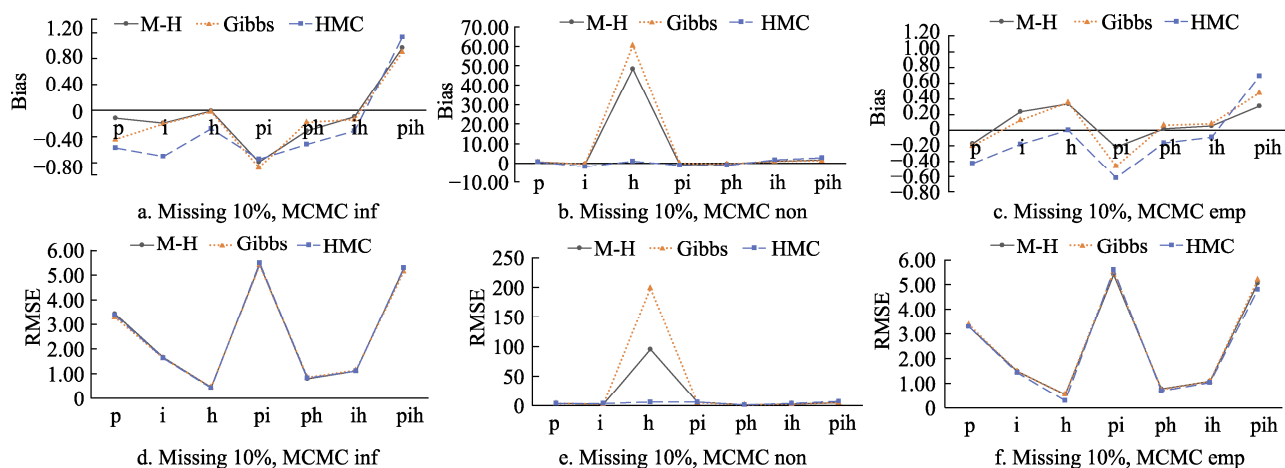
先验信息	抽样方法	估计值	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC inf	M-H	VC	15.502	3.635	0.998	63.292	1.641	2.980	145.044
		Bias	-0.498	-0.365	-0.002	-0.708	-0.359	-0.020	1.044
		RMSE	3.214	1.566	0.419	5.126	0.804	1.142	4.801
	Gibbs	VC	15.598	3.785	0.969	63.422	1.733	3.032	144.715
		Bias	-0.402	-0.215	-0.031	-0.578	-0.267	0.032	0.715
		RMSE	3.420	1.599	0.377	4.992	0.777	1.159	4.752
	HMC	VC	15.391	3.357	0.718	63.386	1.487	2.784	144.717
		Bias	-0.609	-0.643	-0.282	-0.614	-0.513	-0.216	0.717
		RMSE	3.219	1.644	0.446	5.160	0.852	1.177	4.600
MCMC non	M-H	VC	16.648	3.811	56.566	63.578	1.613	3.778	144.613
		Bias	0.648	-0.189	55.566	-0.422	-0.387	0.778	0.613
		RMSE	3.648	2.757	64.149	5.043	1.355	2.027	4.594
	Gibbs	VC	16.506	3.656	159.590	63.320	1.464	3.867	144.886
		Bias	0.506	-0.344	158.590	-0.680	-0.536	0.867	0.886
		RMSE	3.627	2.733	593.552	5.001	1.338	2.164	4.773
	HMC	VC	16.426	1.923	2.083	62.768	0.601	4.750	146.259
		Bias	0.426	-2.077	1.083	-1.232	-1.399	1.750	2.259
		RMSE	3.489	3.251	5.550	5.733	1.767	3.134	5.772
MCMC emp	M-H	VC	15.865	4.212	1.412	63.631	1.977	3.124	144.620
		Bias	-0.135	0.212	0.412	-0.369	-0.023	0.124	0.620
		RMSE	3.136	1.494	0.631	4.828	0.745	1.042	4.685
	Gibbs	VC	15.919	4.115	1.376	63.497	2.025	3.116	144.574
		Bias	-0.081	0.115	0.376	-0.503	0.025	0.116	0.574
		RMSE	3.183	1.511	0.583	4.850	0.716	1.031	4.547
	HMC	VC	15.663	3.817	1.012	63.511	1.823	2.946	144.696
		Bias	-0.337	-0.183	0.012	-0.489	-0.177	-0.054	0.696
		RMSE	3.207	1.383	0.320	5.098	0.688	1.002	4.683

表 5 $p \times i \times h$ 设计方差分量的估计结果(缺失 5%)

先验信息	抽样方法	估计值	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC inf	M-H	VC	15.722	3.767	0.949	63.682	1.664	2.871	144.847
		Bias	-0.278	-0.233	-0.051	-0.318	-0.336	-0.129	0.847
		RMSE	3.210	1.629	0.425	5.351	0.770	1.126	5.046
		VC	15.855	3.642	0.988	63.427	1.756	2.869	144.797
		Bias	-0.145	-0.358	-0.012	-0.573	-0.244	-0.131	0.797
		RMSE	3.427	1.571	0.462	5.270	0.784	1.123	4.853
	HMC	VC	15.285	3.389	0.729	63.232	1.519	2.791	144.992
		Bias	-0.715	-0.611	-0.271	-0.768	-0.481	-0.209	0.992
		RMSE	3.423	1.582	0.445	5.159	0.819	1.130	4.942
MCMC non	M-H	VC	16.639	3.554	91.233	63.441	1.350	3.815	145.196
		Bias	0.639	-0.446	90.233	-0.559	-0.650	0.815	1.196
		RMSE	3.558	2.630	213.273	5.287	1.353	2.104	5.082
		VC	16.495	3.392	62.578	63.518	1.453	3.934	145.250
		Bias	0.495	-0.608	61.578	-0.482	-0.547	0.934	1.250
		RMSE	3.601	2.710	265.197	5.258	1.427	2.130	5.108
	HMC	VC	16.476	2.099	2.016	62.650	0.614	4.707	146.432
		Bias	0.476	-1.901	1.016	-1.350	-1.386	1.707	2.432
		RMSE	3.672	3.243	6.316	6.733	1.782	3.204	6.582
MCMC emp	M-H	VC	15.777	4.215	1.312	63.886	1.996	3.033	144.508
		Bias	-0.223	0.215	0.312	-0.114	-0.004	0.033	0.508
		RMSE	3.299	1.557	0.531	5.277	0.743	1.023	5.091
		VC	15.901	4.127	1.337	63.873	2.059	3.143	144.427
		Bias	-0.099	0.127	0.337	-0.127	0.059	0.143	0.427
		RMSE	3.183	1.418	0.517	5.105	0.754	1.086	4.905
	HMC	VC	15.648	3.811	1.010	63.684	1.825	2.949	144.390
		Bias	-0.352	-0.189	0.010	-0.316	-0.175	-0.051	0.390
		RMSE	3.245	1.423	0.328	5.403	0.718	1.051	4.972

表 6 $p \times i \times h$ 设计方差分量的估计结果(缺失 10%)

先验信息	抽样方法	估计值	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC inf	M-H	VC	15.883	3.801	0.985	63.209	1.686	2.899	144.971
		Bias	-0.117	-0.199	-0.015	-0.791	-0.314	-0.101	0.971
		RMSE	3.421	1.667	0.458	5.427	0.808	1.131	5.252
	Gibbs	VC	15.551	3.792	0.983	63.123	1.813	2.860	144.900
		Bias	-0.449	-0.208	-0.017	-0.877	-0.187	-0.140	0.900
		RMSE	3.281	1.643	0.453	5.393	0.860	1.142	5.148
	HMC	VC	15.428	3.297	0.719	63.256	1.478	2.691	145.132
		Bias	-0.572	-0.703	-0.281	-0.744	-0.522	-0.309	1.132
		RMSE	3.383	1.619	0.423	5.488	0.822	1.109	5.286
MCMC non	M-H	VC	16.583	3.509	49.424	63.045	1.304	3.955	145.767
		Bias	0.583	-0.491	48.424	-0.955	-0.696	0.955	1.767
		RMSE	3.632	2.705	94.872	5.647	1.425	2.297	5.450
	Gibbs	VC	16.515	3.563	61.315	63.673	1.405	3.855	145.220
		Bias	0.515	-0.437	60.315	-0.327	-0.595	0.855	1.220
		RMSE	3.465	2.769	199.282	5.415	1.428	2.219	5.323
	HMC	VC	16.520	2.113	2.070	62.654	0.568	4.713	146.663
		Bias	0.520	-1.887	1.070	-1.346	-1.432	1.713	2.663
		RMSE	3.698	3.280	5.760	6.691	1.786	3.321	6.786
MCMC emp	M-H	VC	15.830	4.230	1.330	63.785	2.020	3.053	144.306
		Bias	-0.170	0.230	0.330	-0.215	0.020	0.053	0.306
		RMSE	3.278	1.480	0.537	5.390	0.771	1.077	5.059
	Gibbs	VC	15.798	4.128	1.355	63.533	2.061	3.075	144.474
		Bias	-0.202	0.128	0.355	-0.467	0.061	0.075	0.474
		RMSE	3.368	1.502	0.553	5.525	0.756	1.086	5.203
	HMC	VC	15.572	3.815	0.995	63.379	1.831	2.908	144.690
		Bias	-0.428	-0.185	-0.005	-0.621	-0.169	-0.092	0.690
		RMSE	3.292	1.444	0.283	5.616	0.677	1.040	4.816

图 2 $p \times i \times h$ 设计方差分量的 Bias 和 RMSE 图(缺失比例 10%)

而在无信息先验条件下(图 2 中的 b 图), M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计偏差非常大, 分别为 48.424 和 60.315, 除去这两个 Bias 值, 三种抽样方法的估计偏差范围为 $(-2, 3)$ 。有信息先验条件下, M-H 算法对 $\sigma^2(p)$ 的估计偏差最接近于 0, M-H 算法和 Gibbs 抽样对 $\sigma^2(i)$ 、 $\sigma^2(h)$ 和 $\sigma^2(ph)$ 的估计偏差比 HMC 算法略小(但相当接近), 三种方法对其余 3 个方差分量的估计偏差接近。在经验性先验条件下, HMC 算法对 $\sigma^2(h)$ 的估计偏差最接近于 0, M-H 算法和 Gibbs 抽样对 $\sigma^2(p)$ 、 $\sigma^2(pi)$ 、 $\sigma^2(ph)$ 和 $\sigma^2(pih)$ 的估计偏差比 HMC 算法略小(但相当接近), 三种方法对 $\sigma^2(i)$ 和 $\sigma^2(ih)$ 的估计偏差接近。

从 RMSE 的结果来看, 在有信息先验和经验性先验条件下, M-H 算法、Gibbs 抽样和 HMC 算法估计方差分量的 RMSE 都基本接近。三种抽样算法估计 $\sigma^2(h)$ 、 $\sigma^2(ph)$ 和 $\sigma^2(ih)$ 的 RMSE 较小, 估计 $\sigma^2(p)$ 、 $\sigma^2(pi)$ 和 $\sigma^2(pih)$ 的 RMSE 较大。在无信息先验条件下, M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 估计的 RMSE 非常大, 分别为 94.872 和 199.282。

总之, 除了在无信息先验条件下, M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计偏差过大, 其他条件下三种抽样方法对方差分量的估计值都在真值上下波动, 这表明在无信息先验下仅 HMC 算法能准确估计方差分量 $\sigma^2(h)$, 其他条件下三种抽样方法的方差分量估计值相对比较接近, 但具体表现不一。

3.4.2 不同 h 侧面水平数下的估计结果

在无信息先验条件下, M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计偏差过大。因此, 增加 h 侧面的水平数, 以探究 M-H 算法和 Gibbs 抽样是否能够更为准确估计 $\sigma^2(h)$ 。在无信息先验条件下, h 侧面的水平数为 3 至 6 时, M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计结果如表 7 和表 8 所示。

表 7 M-H 算法估计 $\sigma^2(h)$ 的结果

缺失比例	估计值	$h=3$	$h=4$	$h=5$	$h=6$
Missing 0%	VC	3.782	2.200	1.965	1.505
	Bias	2.782	1.200	0.965	0.505
	RMSE	6.219	2.598	1.987	1.343
Missing 5%	VC	8.153	2.916	1.789	1.466
	Bias	7.153	1.916	0.789	0.466
	RMSE	12.026	3.397	1.925	1.268
Missing 10%	VC	6.809	2.817	1.782	1.533
	Bias	5.809	1.817	0.782	0.533
	RMSE	9.538	3.310	1.817	1.400

表 8 Gibbs 抽样估计 $\sigma^2(h)$ 的结果

缺失比例	估计值	$h=3$	$h=4$	$h=5$	$h=6$
Missing 0%	VC	8.136	2.965	1.853	1.637
	Bias	7.136	1.965	0.853	0.637
	RMSE	11.989	3.821	1.952	1.512
Missing 5%	VC	5.676	2.769	1.868	1.666
	Bias	4.676	1.769	0.868	0.666
	RMSE	7.781	3.332	1.888	1.543
Missing 10%	VC	6.195	2.526	1.916	1.533
	Bias	5.195	1.526	0.916	0.533
	RMSE	9.044	2.940	1.958	1.405

从表 7 和表 8 可知, 随着 h 侧面水平数的增加, M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计值越来越接近其真值 1, Bias 值和 RMSE 值都逐渐减小。当 h 侧面的水平数为 3 时, M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计偏差依然相对较大。然而, 当 h 侧面的水平数增加为 4 时, M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计偏差较小, 且继续增加为 5 和 6 时, 估计偏差也未增大。 h 侧面的水平数为 4~6 时, 方差分量 Bias 值的区间范围较为稳定, 与有信息先验和经验性先验条件下的区间范围较为接近。当 h 侧

面的水平数增加为 4 或者更多时, M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计偏差较小, 与有信息先验和经验性先验的结果较为接近。因此, 若未来研究者继续采用 M-H 算法和 Gibbs 抽样估计双侧面交叉设计的方差分量时, 则研究中 h 侧面水平数不应低于 4, 以免影响估计结果的准确性。

4 实证研究

4.1 $p \times i$ 设计实证数据

采用 Zung (1965) 编制的抑郁自评量表(Self-Rating Depression Scale, SDS) 对学生进行测量。SDS 包含 20 个反映抑郁主观感受的项目, 其中, 10 项为正向评分题, 另 10 项为反向评分题。选项如“我觉得闷闷不乐, 情绪低沉”“我觉得一天之中早晨最好”等, 按症状出现的频度采用 4 级计分。1 表示“没有或很少时间”, 2 表示“小部分时间”, 3 表示“相当多时间”, 4 表示“绝大部分或全部时间”。

数据来源于中国南方某城市随机取样的 1545 名学生, 对他们进行了间隔各为 6 个月的追踪测试, 共进行了三次纸笔测试。三次测试均存在被试流失, 其原因可能为: (1) 测试当天缺勤; (2) 题目错填或漏填; (3) 转学或休学。剔除未完成三次测试的被试, 最后保留被试 1429 名, 其中, 男性青少年 636 名, 女性青少年 793 名, 平均年龄 14.88 岁(标准差 1.80)。该量表在本研究中的内部一致性表现良好(T1 Cronbach's $\alpha = 0.902$; T2 Cronbach's $\alpha = 0.906$; T3 Cronbach's $\alpha = 0.922$)。

对于随机取样 1429 名青少年, 其中有完整数据的被试为 1288 人, 有缺失数据的被试为 141 人, 占比 9.9%。这种设计为 $p \times i$ 设计, 其中 p 是指青少年(共 1429 名), i 是指抑郁量表的题目(共 20 道题), 取三次测量的题目平均分为题目得分, 即 T1、T2 和 T3 的题目平均得分。对 1429 名被试的数据先重

复 5000 次估计, 将估计的方差分量平均数视为“参数”(真值)(Tong & Brennan, 2007)。因为前人没有该数据的先验信息, 所以用无信息先验 MCMC 方法估计方差分量, 所得结果如表 9 所示。

从表 9 可知, 三种抽样方法的结果都较好, 这与模拟结果的结论一致。相较而言, HMC 整体上优于 M-H 和 Gibbs 方法。

4.2 $p \times i \times h$ 设计实证数据

采用 SDS 量表, 对 1429 名青少年进行间隔一年的重复测量, 这种设计为 $p \times i \times h$ 设计, 其中 p 是指青少年(共 1429 名), i 是指抑郁量表的题目(共 20 道题), h 是指两个时间点(T1 和 T3)的重复测量, 缺失率约为 9.9%。与单侧面交叉设计 $p \times i$ 设计一样, 对 1429 名被试的数据先重复 5000 次估计, 将估计的方差分量平均数视为“参数”(真值)(Tong & Brennan, 2007), 且因为前人没有该数据的先验信息, 所以用无信息先验 MCMC 方法估计方差分量, 所得结果如表 10 所示。

从表 10 可知, 较大的偏差主要表现在 M-H 和 Gibbs 的 $\sigma^2(h)$ 上, 但 HMC 方法的 $\sigma^2(h)$ 仍然较小, 这与模拟结果的结论一致。

5 综合讨论

5.1 三种 MCMC 抽样方法的比较

在单侧面交叉设计 $p \times i$ 模拟研究中, M-H 算法、Gibbs 抽样和 HMC 算法在不同条件下都能较为准确地估计方差分量。在双侧面交叉设计 $p \times i \times h$ 模拟研究中, 有信息先验和经验性先验条件下, M-H 算法、Gibbs 抽样和 HMC 算法的估计准确性相当; 而在无信息先验条件下, HMC 算法对方差分量的估计准确性更高。M-H 算法和 Gibbs 算法在无信息先验条件下估计双侧面交叉设计的方差分量 $\sigma^2(h)$ 时, 估计偏差都过大。无论是在单侧面交叉

表 9 $p \times i$ 设计方差分量的估计结果(缺失 9.9%)

先验信息	抽样方法	估计值	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
MCMC non	M-H	VC	4.005	17.800	65.102
		Bias	0.005	1.800	1.102
		RMSE	1.091	6.306	2.189
	Gibbs	VC	3.999	17.089	64.199
		Bias	-0.001	1.089	0.199
		RMSE	1.189	2.501	2.236
	HMC	VC	3.999	16.704	64.177
		Bias	-0.001	0.704	0.177
		RMSE	1.255	2.788	2.202

表 10 $p \times i \times h$ 设计方差分量的估计结果(缺失 9.9%)

先验信息	抽样方法	估计值	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC non	M-H	VC	16.983	3.559	49.404	63.245	1.388	3.988	145.700
		Bias	0.983	-0.441	48.404	-0.755	-0.612	0.988	1.700
		RMSE	3.639	2.745	92.002	5.007	1.400	2.207	5.459
	Gibbs	VC	16.515	3.563	61.309	63.673	1.405	3.889	145.288
		Bias	0.515	-0.437	60.309	-0.327	-0.595	0.889	1.288
		RMSE	3.465	2.769	190.202	5.415	1.428	2.208	5.309
	HMC	VC	16.529	2.119	2.071	62.600	0.587	3.753	146.612
		Bias	0.529	-1.881	1.071	-1.400	-1.413	0.753	2.612
		RMSE	3.690	3.289	5.761	6.608	1.789	3.308	6.7822

设计中,还是在双侧面交叉设计中,M-H 算法和 Gibbs 抽样对方差分量的估计结果均较为一致。甚至于,在双侧面交叉设计无信息先验条件下,两者都对 $\sigma^2(h)$ 的估计呈现出高度一致的偏差。M-H 算法和 Gibbs 抽样对方差分量估计表现的一致性,与两种抽样方法的原理和联系紧密相关。

在实证研究中,不论是 $p \times i$ 设计实证数据,还是 $p \times i \times h$ 设计实证数据,其结果与模拟研究的结论几乎完全一致。在 $p \times i$ 设计实证数据中,M-H 方法、Gibbs 方法和 HMC 方法的结果都较好,但是 $p \times i \times h$ 设计实证数据中,M-H 方法和 Gibbs 方法在估计 $\sigma^2(h)$ 时,出现了较大的偏差且 RMSE 也较大,而 HMC 方法的 $\sigma^2(h)$ 偏差却依然较小。

HMC 算法是将汉密尔顿系统的物理特性和 Leapfrog 技术相结合完成抽样的一种抽样方法,属于全局更新的算法;运算效率较高,能处理复杂模型的求解问题(Lopilato et al., 2015; Nugroho & Morimoto, 2015; Pakman & Paninski, 2014)。HMC 算法显示出其优势特征:双侧面交叉设计中,在无信息先验条件下,HMC 算法能够准确估计 $\sigma^2(h)$,而 M-H 算法和 Gibbs 抽样的估计偏差都较大。这表明,HMC 算法具有较强的稳健性。

5.2 双侧面交叉设计 h 侧面水平数的拐点

双侧面交叉设计中无信息先验条件下的估计结果可能与 h 侧面的水平数为 2 有关。在后续的研究中,随着 h 侧面水平数的增加,M-H 算法和 Gibbs 抽样的估计偏差逐渐减小,尤其是增加到 4 之后,其估计结果较为接近有信息先验和经验性先验条件下的估计表现。这表明,在无信息先验条件下,M-H 算法和 Gibbs 抽样对 $\sigma^2(h)$ 的估计表现较差,与 h 侧面的水平数较小有关。

Moore (2010)研究了前人研究中使用过测量侧

面的水平数,并用考生相对于题目的比例(Ratio of persons to items)表示侧面水平数之比,得出其范围为 0.13 到 120。根据以往研究中的侧面水平数之比,Moore (2010)确定了相应研究中的测量侧面数量,并发现测量侧面水平数对方差分量的估计结果产生一定的影响。当 h 侧面的水平数为 2 时,M-H 算法和 Gibbs 抽样的估计不理想,而 HMC 算法理想。但是,当 h 侧面的水平数为 4 及以上时,M-H 算法和 Gibbs 抽样与 HMC 算法估计相当。这进一步表明,HMC 算法具有较强的稳健性,能够克服 h 侧面的水平数带来的影响。若 M-H 算法和 Gibbs 抽样要与 HMC 算法估计性能相当,则需满足双侧面交叉设计 h 侧面水平数大于或等于 4。这表明,双侧面交叉设计在 h 侧面水平数取 4 时出现拐点。

HMC 算法在多数情况下与 M-H 算法、Gibbs 抽样的估计表现一致,都能较为准确地估计单侧面交叉设计和双侧面交叉设计的方差分量。但是,M-H 算法和 Gibbs 抽样容易受到 h 侧面的水平数带来的影响,表现出不够稳健。总之,与 Gibbs 抽样和 M-H 算法相比,HMC 算法整体上表现得更优,值得推广。

5.3 先验信息和缺失比例对方差分量的影响

先验信息会影响方差分量的估计,而缺失比例对方差分量的估计影响很小。在单侧面交叉设计中,三种先验分布条件下方差分量的估计表现差异较小,且有信息先验和经验性先验条件下方差分量的估计准确性都较高。在双侧面交叉设计中,三种先验信息对方差分量的估计结果影响较为明显,也是有信息先验和经验性先验条件下方差分量的估计准确性更好,这与前人研究结果一致(Lopilato et al., 2015)。

在单侧面交叉设计和双侧面交叉设计中,三

种缺失比例条件下, 方差分量的估计表现差异都很小。这也进一步验证了 MCMC 方法在 GT 中能够较好地处理缺失数据, MCMC 方法处理 GT 的缺失数据表现出较强的稳健性。MCMC 方法能够非常有效地处理 GT 分析中的缺失数据, 表现出较强的优势, 这是其它方差分量估计方法所不能比拟的。

从模拟研究运行时间来看, 与单侧面交叉设计相比, 双侧面交叉设计耗费的时间更久, 说明研究设计越复杂, MCMC 方法的计算时间越长。与 M-H 算法和 Gibbs 抽样相比, HMC 算法的运算时间稍长, 这可能与算法的复杂性有关, 但如今的计算机运行速度都相对较快, 少量的运行时间增长, 并不会严重影响估计的结果。

5.4 局限与展望

其一, 仅涉及了 GT 研究中的两种常见测验设计——单侧面交叉设计和双侧面交叉设计, 未来研究可以继续探讨不同抽样方法在其他的测验设计对方差分量的估计表现, 如嵌套设计和混合设计等(Vispoel et al., 2023)。其二, 缺失数据在实际的测量过程中较为常见, 仅探讨了完全随机缺失对方差分量估计准确性的影响, 未来研究可增加对其他缺失机制的考察, 如随机缺失和非随机缺失等(Van Hooijdonk et al., 2022)。其三, 模拟研究固定学生人数和题量, 未探讨改变样本量对方差分量估计的影响, 且 $h \geq 4$ 是基于 $p \times i$ 和 $p \times i \times h$ 两种设计之下所获得的, 未来研究可探讨改变学生人数和题量对方差分量估计的影响, 也可探讨基于更多设计之下 h 侧面样本量对方差分量估计的影响。

6 结论

(1)少量数据缺失对 MCMC 三种抽样方法几乎无影响, 表明 MCMC 对于缺失数据估计具有较强的稳健性。

(2)整体上, 对于方差分量估计而言, 有信息先验估计和经验性先验估计较无信息先验估计准确性更高。

(3)仅 HMC 算法表现出“跨设计性”, 在两种设计之下的估计准确性都较高, 但 M-H 算法和 Gibbs 抽样却受限于 $\sigma^2(h)$, 估计准确性都不高。

(4)若 M-H 算法和 Gibbs 抽样需要克服 $\sigma^2(h)$ 估计准确性不高的局限性, 则需规定 $h \geq 4$; 未来研究者在使用 MCMC 方法时, HMC 算法更值得推荐。

参考文献

- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Brennan, R. L., Kim, S. Y., & Lee, W. C. (2022). Extended multivariate generalizability theory with complex design structures. *Educational and Psychological Measurement*, 82(4) 617–642.
- Bourabee, N., & Sanzserna, J. M. (2017). Randomized Hamiltonian Monte Carlo. *Annals of Applied Probability*, 27(4), 2159–2194.
- Chaudhuri, S., Mondal, D., & Yin, T. (2017). Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 79(1), 293–320.
- Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal*, 59(6), 1122–1156.
- Enders, C. K., Fairchild, A. J., & MacKinnon, D. P. (2013). A Bayesian approach for estimating mediation effects with missing data. *Multivariate Behavioral Research*, 48(3), 340–369.
- Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593–1623.
- Huang, C. (2017). *Bayesian statistics and its implementation in R*. Beijing: Tsinghua University Press.
- [黄长全. (2017). *贝叶斯统计及其R实现*. 北京: 清华大学出版社.]
- Jackson, D. J. R., Michaelides, G., Dewberry, C., & Kim, Y. J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology*, 101(7), 976–994.
- Jiang, Z., & Skorupski, W. (2018). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior research methods*, 50, 2193–2214.
- Li, G. (2016). Influence factors of estimating of variance components and their variability for generalizability study in generalizability Theory. *Psychological Exploration*, 36(5), 458–463.
- [黎光明. (2016). 概化理论 G 研究方差分量及其变异量估计影响因素. *心理学探新*, 36(5), 458–463.]
- Li, G. (2023). How many students and items are optimal for teaching level evaluation of college teachers? Evidence from generalizability theory and Lagrange multiplier. *Sustainability*, 15(1), 2.
- Li, G., & Qin, Y. (2022). A new method for estimating the optimal sample size in generalizability theory based on evolutionary algorithm: Comparisons with three traditional methods. *Acta Psychologica Sinica*, 54(10), 1262–1276.
- [黎光明, 秦越. (2022). 一种基于进化算法的概化理论最佳样本量估计新方法: 兼与三种传统方法比较. *心理学报*, 54(10), 1262–1276.]
- Li, G., Wang, X., & Pan, Y. (2022). Comparing of different methods for estimating variance components and their variabilities in generalizability theory based on two-facet designs. *Statistics and Decision*, 38(3), 50–54.
- [黎光明, 王幸君, 潘语熙. (2022). 概化理论两侧面设计方差分量及其变异量估计方法比较. *统计与决策*, 38(3), 50–54.]
- Li, G., & Wu, Y. (2023). Application of the Bookmark method:

- Setting standard for the ninth-grade mathematics achievement test in China. *Current Psychology*, 42, 28941–28952.
- Liao, R. J. T. (2023). The use of generalizability theory in investigating the score dependability of classroom-based L2 reading assessment. *Language Testing*, 40(1), 86–106.
- Liu, J., & Xia, Q. (2016). *Bayesian statistical method based on MCMC algorithm*. Beijing: Science Press.
- [刘金山, 夏强. (2016). *基于MCMC算法的贝叶斯统计方法*. 北京: 科学出版社.]
- Lopilato, A. C., Carter, N., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management*, 41(2), 692–717.
- Mao, X., Shin, D., & Brennan, R. (2005). *Estimating the variability of the estimated variance components and related statistics using the MCMC procedure: An exploratory study*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.
- Medvedev, O. N., & Siegert, R. J. (2022). Generalizability theory. In O. N. Medvedev, C. U. Krägeloh, R. J. Siegert, & N. N. Singh (Eds.), *Handbook of assessment in mindfulness research* (pp. 1–12). Cham: Springer International Publishing.
- Medvedeva, A., Galli, G., Petrovskaya, A., & Medvedev O. N. (2022). A novel application of generalizability theory to evaluate the reliability of the recognition memory test. *Current Psychology*, 42(20), 17020–17032.
- Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3), 339–348.
- Moore, J. L. (2010). *Estimating standard errors of estimated variance components in generalizability theory using bootstrap procedures*. (Doctor of Philosophy Doctor), University of Iowa.
- Neal, R. M. (2010). *MCMC using Hamiltonian dynamics*. London: Chapman and Hall.
- Nugroho, D. B., & Morimoto, T. (2015). Estimation of realized stochastic volatility models using Hamiltonian Monte Carlo-Based methods. *Computational Statistics*, 30(2), 491–516.
- Pakman, A., & Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational & Graphical Statistics*, 23(2), 518–542.
- Tong, Y., & Brennan, R. L. (2007). Bootstrap estimates of standard errors in generalizability theory. *Educational and Psychological Measurement*, 67(5), 804–817.
- Van Hooijdonk, M., Mainhard, T., Kroesbergen, E. H., & van Tartwijk, J. (2022). Examining the assessment of creativity with generalizability theory: An analysis of creative problem solving assessment tasks. *Thinking Skills and Creativity*, 43, 100994.
- Vispoel, W. P., Hong, H., Lee, H., & Jorgensen, T. D. (2023). Analyzing complete generalizability theory designs using structural equation models. *Applied Measurement in Education*, 36(4), 372–393.
- Wang, X., Li, G., Zhang, M., Jiang, X., Liang, Z., & Chu, X. (2016). A generalizability analysis of teaching level evaluation for college teachers. *Journal of Psychological Science*, 39(1), 90–96.
- [王幸君, 黎光明, 张敏强, 蒋欣, 梁正妍, 楚肖依. (2016). 高校教师教学水平评价的概化理论分析. *心理科学*, 39(1), 90–96.]
- Yang, Z., & Chang, L. (2003). *Generalizability theory and its applications*. Beijing, China: Educational Science Publishing House.
- [杨志明, 张雷. (2003). *测评的概化理论及其应用*. 北京: 教育科学出版社.]
- Zhang, C., Shahbaba, B., & Zhao, H. (2017). Hamiltonian Monte Carlo acceleration using surrogate functions with random bases. *Statistics and Computing*, 27(6), 1473–1490.
- Zhao, X., Ren, Z., & Chen, A. (2018). Analysis and comparative study on the quality of mathematics, arts, and science test papers in the college entrance examination based on multivariate generalizability theory. *Mathematics Bulletin*, 57(1), 25–30.
- [赵轩, 任子朝, 陈昂. (2018). 基于多元概化理论的高考数学文理科试卷质量分析与对比研究. *数学通报*, 57(1), 25–30.]
- Zheng, X., Liu, Y., & Li, G. (2025). Using multivariate generalizability theory to test construct validity of the in-basket test and optimize evaluation plan. *Journal of Jiangxi Normal University (Natural Science Edition)*, 49(6), 623–634.
- [郑雪琴, 刘耀中, 黎光明. (2025). 基于多元概化理论的公文筐测验构思效度检验与测评方案优化. *江西师范大学学报(自然科学版)*, 49(6), 623–634.]
- Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry*, 12(1), 63–70.

Comparing the accuracy of three sampling algorithms of MCMC method for estimating variance components in generalizability theory

LI Guangming

(School of Psychology, Center for Studies of Psychological Application, South China Normal University, Guangzhou 510631, China)

Abstract

Generalizability theory (GT), a modern psychometric framework, has been widely applied in practice. Its efficacy relies on the accurate estimation of variance components. While variance components are traditionally estimated using methods such as analysis of variance (ANOVA), Bootstrap, Jackknife, MINQUE, and restricted maximum likelihood (REML method), Bayesian Markov chain Monte Carlo (MCMC) estimation has demonstrated advantages. However, the

performance of MCMC depends on the chosen sampling algorithms. Local update algorithms, such as Metropolis-Hastings (M-H) algorithm and Gibbs sampling, are prone to random walk behaviors and slow convergence, particularly when conditional distributions are complex or parameters are highly correlated. Although the Hamiltonian Monte Carlo (HMC) algorithm effectively suppresses random walk behavior and handles high-dimensional, complex models efficiently, its application to the estimation of variance components within GT remains unexamined. To date, no published studies have compared M-H algorithm, Gibbs sampling, and HMC algorithm in this context.

This study evaluated and compared the estimation accuracy of three sampling algorithms of MCMC method across two common GT research designs: the one-faceted crossed design ($p \times i$) and the two-faceted crossed design ($p \times i \times h$). Using simulation methodologies, we manipulated three independent variables: (1) sampling algorithms (M-H, Gibbs, and HMC), (2) prior distributions (informative, non-informative, and empirical), and (3) missing data ratios (0%, 5%, and 10%). The dependent variables were the estimated variance components, Bias, and Root Mean Square Error (RMSE).

In simulation study, we generated data using R software, with 1000 batches of data generated for the $p \times i$ and the $p \times i \times h$. For the generated data, R software implemented the estimation of variance components for M-H algorithm, Gibbs sampling, and HMC algorithm by calling the corresponding software package. In empirical research, a batch of data containing 1429 participants was used to validate the conclusions drawn from the simulation study, and it was found that the empirical research results were almost consistent with the simulation research results.

The research results indicated that: (1) low ratios of missing data (up to 10%) yielded negligible effects across all three algorithms, demonstrating the robust handling of missingness inherent to MCMC; (2) across all conditions, informative and empirical priors yielded higher estimation accuracy than non-informative priors; (3) HMC algorithm uniquely exhibited “cross-design” stability, maintaining high accuracy across both the $p \times i$ and the $p \times i \times h$, whereas M-H algorithm and Gibbs sampling demonstrated structural limitations and lower accuracy; and (4) for M-H algorithm and Gibbs sampling to achieve acceptable accuracy in higher-dimensional spaces, a minimum of four levels for the h facet ($h \geq 4$) was required. Consequently, we highly recommend that future GT researchers utilizing Bayesian MCMC frameworks adopt the HMC algorithm.

Keywords generalizability theory, estimating variance components, MCMC method, M-H algorithm, Gibbs sampling, HMC algorithm