文章编号:1000-5862(2020)03-0282-10

应用 Stan 软件包实现 IRT 模型的贝叶斯参数估计

刘思杨,蔡 艳*
(江西师范大学心理学院,江西 南昌 330022)

摘要: Stan 是一个新的用于估计指定统计模型的概率编程语言,它使用了强大而高效的汉密尔顿蒙特卡罗(Hamiltonian Monte Carlo,HMC)抽样算法,相比较传统的 Gibbs 抽样和 Metropolis 算法具有显著的效率提升. R 软件包"rstan"链接了 R 与 Stan 2 个软件,使得 Stan 可以借助 R 的计算环境运行. 首先,该文通过 3 参数 Logistic(3PL)模型代码介绍了 Stan 的程序语言;其次,该文使用 Stan 计算 2 个评估模型-数据拟合的全新指标 WAIC 和 LOO,为应用 Stan 进行 IRT 模型相关研究提供了有效的参考工具;最后,该文还采用了 2 个真实数据分别考察了 Stan 在单维 IRT 模型和多维 IRT 模型参数估计中的运行表现. 研究结果表明:采用一个新的贝叶斯统计软件 Stan,通过 2 个实证研究验证了该方法的有效性与可行性,为国内学者应用 Stan 进行 IRT 模型相关研究提供了有效的参考资料.

关键词: 项目反应理论;汉密尔顿蒙特卡罗算法;贝叶斯估计;Stan
中图分类号: B 841 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2020.03.12

0 引言

近 30 年来,随着计算机技术的不断发展和计算能力不断提高,以及马尔科夫链-蒙托卡罗(Markov Chain Monto Carlo,MCMC)模拟技术不断成熟,基于贝叶斯统计技术的项目反应理论(IRT)建模越来越受欢迎[1]. 与此同时,基于 MCMC 算法的贝叶斯统计软件的出现以及不断迭代更新,将异常复杂的后验分布计算变得简单并且易操作,大大降低了研究者采用贝叶斯统计方法进行应用研究的门槛,进一步促进了贝叶斯方法在其他领域中的推广和应用[2].

目前,有多个统计软件可用来实现 MCMC 算法,包括 WinBUGS[3]、OpenBUGS[4]、JAGS[5]、PROC MCMC in SAS 和 R 软件包"mcmcpack"[6]等. 当然,这些统计软件均是基于 Gibbs 抽样算法[7]和 Metropolis 算法[8]开发的. 尽管这 2 种 MCMC 算法被广泛使用,但是这些算法仍然具有一些限制,具体表现为这 2 种算法在模型后验参数空间中的搜索效率较低,导致模型收敛所需的计算时间较长[9].

Stan[10]是一个全新的贝叶斯软件,是对汉密尔顿蒙特卡罗算法(HMC)[11]的扩展应用. 与传统的 Gibbs 抽样算法和 Metropolis 算法相比,HMC 算法的运行速度提升较大,能更有效地探索后验参数空间. 具体表现为 HMC 算法通过将每个模型参数与动态变量相配对,并根据当前抽样的参数值的后验密度,完成对目标变量分布空间的探索抽样,从而有效地解决了 Metropolis 算法中随机游走(random walk)抽样效率较低的问题[12]. 因此,基于 HMC 算法的 Stan 被认为比传统的贝叶斯软件更加高效[13]. 如在 Stan 中可能只需要 1 000 次迭代就可得到较好的收敛效果,而在传统的 BUGS 软件中达到同样的收敛效果则可能需要 100 000 次[14]. Stan 还允许使用不恰当的先验分布,对于参数先验分布的设定具有较好的容错性,相比较 WinBUGS 和 OpenBUGS,这是其另外一个优势. 此外,Stan 具有较好的软件和系统的兼容性,不仅支持 R、Python、Matlab、Stata 等软件的接口,还支持 Windows、Mac4、Linux 等系统[1].

多维项目反应理论(Multidimensional Item Response Theory,MIRT)是项目反应理论(Item Response Theory,IRT)的一个重要发展方向,涂冬波等[15]使用 MCMC 算法对 MIRT 模型进行了参数估计的研究. 而国外有些学者系统地介绍了使用 WinBUGS 估计 IRT 模型的代码[16],以及近几年一些研究者使用 SAS PROC MCMC 做了 IRT 模型参数估计研究[17]. 如前所述,虽然 Stan 相比较传统的 MCMC

收稿日期:2019-12-16
基金项目:国家自然科学基金(31760288)资助项目.
通信作者:蔡 艳(1979-),女,江西宜春人,教授,博士生导师,主要从事心理统计与测量研究. E-mail:cy1979123@aliyun.com

统计软件具有较强的优势,但目前国内学者对于 Stan 的了解还非常有限,截至目前为止国内介绍和使用 Stan 的相关文献较少[18]. 甚至在 MIRT 领域中还没有介绍使用 Stan 进行相关研究的文献,这不利于国内 MIRT 的发展. 因此,系统介绍 Stan 的 IRT 模型参数估计代码是十分有必要的,这不仅拓展了 MCMC 软件 Stan 的应用领域范围,而且也丰富了 MIRT 应用研究的工具类别. 本文聚焦于多维项目反应理论模型的参数估计代码的介绍,但为了便于读者理解也会相应呈现单维 IRT 模型参数估计代码并进行解释. 本文主要介绍 R 软件包“rstan”的使用,它链接了 R 与 Stan 2 个软件,使得 Stan 可以借助 R 的计算环境运行. 因此,越来越受到心理测量学者们和应用研究人员的欢迎[19].

1　Stan 代码模块

如前所述,本文采用 R 包“rstan”[20] 在 R 的计算环境中调用 Stan 软件来估计指定的统计模型. 为了使用 Stan 的功能,必须将 Stan 程序指定为以“.stan”作为后缀的单独 Stan 文件,或者作为 R 环境中的对象. 考虑到运行效率,本文采用前一种方式. 另外,在 R 工作目录文件夹中为目标统计模型保存一个单独的 Stan 文件. Stan 运行需要 2 种代码: 第 1 种为 Stan 代码,格式为 C + + 程序语言,由若干提前定义的按顺序排列的模块(Block)组成,这些模块包括数据(Data)、参数(Parameters)、参数转换(Transformed Parameters)、模型(Model)、预测值(Generated Quantities)等; 第 2 种为 R 程序语言格式的 R 代码,通过 R 代码调用已定义的 Stan 文件来运行 Stan. “rstan”软件包的具体安装与调试读者可以参见 Stan 官方网站(http://mc-stan.org/interfaces/). 本文所使用的 R 版本为 3.6.1,R 包“rstan”的版本是发布于 2019 年 7 月 9 日的 v2.19.2.

1.1　数据模块

数据模块(Data Block)是 Stan 程序中必需的模块之一,定义了贝叶斯统计模型中导入数据的各项信息,如观察变量(包括被试和项目等)的数量、类型、取值范围等. 数据模块和 Stan 中的其他模块一样均包含一对“{}”,Stan 代码写在“{}”中,写作 Data{}; “//”之后的语句 Stan 不会识别运行,用于添加注释,对代码进行简要说明; Stan 提供了“〈〉”,供用户定义数据的边界,其中“lower =”和“upper =”分别定义了目标对象的下界和上界; “[]”用来定义目标变量的长度. 值得注意的是,Stan 中每一行代码均需要使用“;”结尾. 用户定义的数据边界和长度要与数据情况相吻合,否则 Stan 会报错误信息,并标注出代码错误处,供用户参考并修改. Stan 数据模块代码为

```
Data {
    int〈lower = 1〉J; //number of subjects
    int〈lower = 1〉K; //number of questions
    int〈lower = 1〉N; //number of observations
    int〈lower = 1,upper = J〉jj [N];
    //subject for observation n
    int〈lower = 1,upper = K〉kk [N];
    //question for observation n
    int〈lower = 0,upper = 1〉Y [N];
    //correctness for observation n
}.
```

从上面的代码中可以看到数据模块中定义的数据信息: “int”表示定义的对象是整数格式,对于非整数连续的数据对象可以使用“real”定义. J 表示作答者的人数,K 表示题目的个数,N 表示所有被试的作答个数(JK). 需要特别说明的是,在 Stan 中运算效率最高的数据类型为向量,因此原始作答矩阵在数据模块中被转化成作答向量,而 jj 数组和 kk 数组分别存储了原始作答数据矩阵中所有列向量的行和所有行向量的列. 例如,在 1 000 个被试 10 道题的作答矩阵转化为作答向量后,jj 为一个 1 ~ 1 000(间隔为 1)且循环 10 次的长度为 1 000 × 10 的数组; kk 为一个 1 ~ 10(间隔为 1)且循环 1 000 次的长度为 10 × 1 000的数组. Y 是一个长度为 N 的观察数据向量.

1.2　参数模块

紧接数据模块的是参数模块(Parameters Block),它定义了模型参数的类型和取值范围,以及指定一些模型参数的先验超参数. 需要特别说明的是,Stan 用户应该在参数模块中定义所有模型中出现的参数或超参数,否则 Stan 会报告警告信息并停止运行. 因此,在编写参数模块时用户应该对模型的参数和超参数有较为详尽地了解.

在 IRT 模型中,3 参数 Logistic 模型(3PL)[21] 参数包括被试的潜在特质参数(θ_i)和 3 个项目参数(区分度参数(a_j)、难度参数(b_j)和猜测度参数(c_j)). 3PL 的模型表达式为

$$p_{ij}(u_{ij}=1 \mid \theta_i, a_j, b_j, c_j) = c_j + (1-c_j)/(1+\exp(-1.7a_j(\theta_i - b_j)));$$

其中 p_{ij} 表示被试 i 在项目 j 上正确作答的概率; u_{ij} 表

示被试 i 在项目 j 上的作答情况,取值为 0 或 1. 另外,通过将区分度参数定义为 1 以及猜测参数定义为 0,3PL 可以简便地转化为 1PL 或 2PL. 因此,示例选择 3PL 代码作讲解说明,代码为

```
Parameters {
    vector〈lower = -3,upper = 3〉[J] theta;
    //ability for person j
    vector〈lower = 0,upper = 3〉[K] alpha;
    //discrimination of item k
    vector[K] beta; //difficulty for item k
    vector〈lower = 0,upper = 0.5〉[K] gamma;
    //item pseudo-guessing for item k
    real mu_beta; //mean of item difficulty
    real〈lower = 0〉sigma_beta;
    //scale of item difficulty
    real〈lower = 0〉sigma_alpha;
    //scale of discrimination
}.
```

在该 Stan 代码中,3PL 的所有模型参数均在参数模块中定义了其类型和范围. 例如,被试的潜在特质参数 theta 被定义为一个取值范围为 [-3,3]、长度为 J 的数组. 关于超参数的定义,可以看到代码中难度参数 beta 的未知均值变量 mu_beta 被定义为实数(real)类型,表示模型中定义的 mu_beta 是一个连续的变量;区分度参数 alpha 的未知标准差 sigma_alpha 和难度参数 beta 的位置标准差 sigma_beta 均被定义为一取值大于 0 的连续变量. 参数模块的功能只是定义所有模型中出现的参数及超参数,模型参数和超参数的先验分布等信息需要在模型模块(Model block)中被定义.

1.3 模型模块

模型模块(Model block)是 Stan 程序中最基本的模块,用于定义模型中出现的所有参数的先验分布以及观察变量的概率分布. 即参数模块中定义的所有参数及超参数的先验分布均必须在模型模块中给出. 需要说明的是,若某个参数的先验分布未在模型模块中给出,则 Stan 程序会默认该参数服从均匀分布. 例如,参数模块中定义被试潜在特质参数 theta 的取值范围是 [-3,3],若在模型模块中未定义 theta 服从标准正态分布,则 Stan 会默认 theta 服从最小值为 -3、最大值为 3 的均匀分布. 3PL 的模型模块 Stan 代码为

```
Model {
    theta ~ std_normal();
    alpha ~ lognormal(0,sigma_alpha);
    beta ~ normal(mu_beta,sigma_beta);
    gamma ~ beta(2,5);
    mu_beta ~ cauchy(0,5);
    sigma_beta ~ cauchy(0,5);
    sigma_alpha ~ cauchy(0,5);
    for (n in 1:N) {
        real p;
        p = inv_logit(1.7* alpha[kk[n]] *
(theta[jj[n]] - beta[kk[n]]));
        Y[n] ~ bernoulli(gamma[kk[n]] + (1 -
gamma[kk[n]]) * p);
    }
}.
```

在以上代码中,分为 2 部分. 第 1 部分定义了所有模型参数的先验分布;第 2 部分通过指定观察变量的概率分布定义似然函数. 在 3PL 模型中,theta 服从均值为 0、标准差为 1 的标准正态分布;alpha 参数服从对数正态分布;beta 参数服从标准正态分布;gamma 参数服从 beta 分布. 值得注意的是,由于 sigma_beta 和 sigma_alpha 在 parameters 模块中被限制为非负,所以这个作为超参数先验的柯西分布实际上是半个柯西分布. 与任何其他贝叶斯软件类似,先验可以是信息丰富的、弱信息丰富的或非信息丰富的. 因此,先验的选择应该反映对这些参数的理解. 在指定完所有参数的先验和超先验之后,通过定义统计模型来指定观察变量的概率分布进而得到似然函数.

1.4 预测值模块

预测值模块(Generated Quantities Block)是 Stan 代码的一个可选的模块,通常用于需要计算新变量并获得其相应的后验分布. 在 IRT 领域中,预测值模块可用于计算基于模型的对数似然,进而计算 2 个用于模型比较和选择的模型拟合指标:LOO(Leave-one-out cross-validation)[23] 和 WAIC(Widely available information criterion)[23]. 若研究者对模型比较和选择不感兴趣,则可以删除 IRT 模型 Stan 代码中的预测值模块. 像 BUGS 等传统的统计软件,通常使用偏差信息指标(DIC)[24] 作为模型比较的指标. 而 Stan 并不计算 DIC,相反地,它将 WAIC 和 LOO 这 2 个指标用于模型比较和选择,因为它们完全是基于贝叶斯理论构建的指标,在理论上优于传统的基于信息的模型选择指标,如赤池信息指标(AIC)[25]、贝叶斯信息指标(BIC)[26] 和 DIC. 在 IRT 模型选择的背景下,Luo Yong 等[27] 研究了 WAIC 和 LOO 在 2 级计分的 IRT 模型上的性能表现,发现它

们优于更传统的方法,如似然比检验、AIC、BIC 和 DIC 等. WAIC 和 LOO 的计算可以是密集的、非平凡的,通常必须使用近似方法. 需要特别说明的是 R 包"loo"已经开发出来,可以与"rstan"包结合使用计算 WAIC 和 LOO 2 个指标. 在下文中,将演示如何使用"rstan"和"loo"2 个 R 包比较和选择合适的 IRT 模型. 预测值模块的 Stan 代码为

```
Generated quantities {
    vector[N] log_lik;
    for(n in 1:N) {
        real pY;
        pY = inv_logit(1.7* alpha[kk[n]] *
(theta[jj[n]] - beta[kk[n]]));
    log_lik[n] = bernoulli_lpmf(Y[n] | gamma[kk[n]] +
(1 - gamma[kk[n]]) * pY);
    }
}.
```

1.5　参数转换模块

参数转换模块(Transformed Parameters Block)和预测值模块一样,也是可选择的模块,当需要转换参数模块中指定的某些参数时使用. 参数转换模块通常用来限制和约束在参数模块中已经定义好的参数或利用已有的参数定义新的模型参数. 例如,在下列参数转换模块代码中,定义了一个新的变量 beta_j,将参数模块中的 beta 参数的第 1 项固定为 0 的并命名为新的参数 beta_j.

```
Transformed parameters {
    vector[K] beta_j;
    beta_j[1] = 0;
    for(i in 2:K) {
        beta_j[i] = beta[i];
    }
}.
```

2　示例研究

本文将使用真实数据来演示如何使用 Stan 分析 IRT 中单维和多维的 2 级计分的项目反应数据. 在真实数据研究中包括 3 个相互竞争的 2 级计分的单维 IRT 模型(1PL、2PL、3PL)以及 2 个多维 IRT 模型(M2PL、M3PL)之间的模型选择. 研究者需事先将 Stan 相关的 R 包"rstan"和"loo"安装好,具体 R 包的安装方法与步骤读者可以参见 R 官方网址(https://www.r-project.org).

2.1　单维真实数据分析

单维真实数据分析采用的数据为心理类型量表(Myers-Briggs Type Indicator; MBTI),该量表为心理学家荣格在 1962 年编制的人格自陈量表,包含 4 个维度: 外向-内向、感觉-直觉、思维-情感、判断-观察; 更多的数据详细信息读者可参见文献[28]. 本文从中抽取 1 个维度: 外向-内向,共 21 个项目,被试 773 人.

3 个 IRT 单维模型(1PL、2PL、3PL)会被用于真实数据的模型拟合比较. 设定储存为".csv"格式的数据文件与 Stan 文件以及 R 文件已经保存在 R 软件的工作路径目录下. 为了便于分析,在本例中将 3 个模型的 Stan 文件名称分别定义为"1PL.stan"、"2PL.stan"、"3PL.stan"; 然后使用 R 代码分别调用 3 个 stan 文件进行 MCMC 运算. "3PL.stan"的 R 代码为

```
library("rstan")
library("loo")
Data <- as.matrix(read.csv("C:/Users/Documents/R/Y.csv")) # 读取数据矩阵
J <- nrow(Data) # 从数据矩阵中读取人数
K <- ncol(Data) # 从数据矩阵中读取题数
Y <- as.vector(Data) # 数据矩阵转换为向量
N <- J* K# 观察数据的个数
jj <- rep(1,K) %x% c(1:J)
kk <- c(1:K) %x% rep(1,J)
data_3pl <- list(J = J,K = K,N = N,jj = jj,
kk = kk,Y = Y)
fit_3pl <- stan(file = "3PL.stan",data =
data_3pl,chains = 1,iter = 5 000,warmup = 2 500,
cores = 4).
```

在该示例代码中,首先载入 2 个 R 包"rstan"和"loo"; 然后读入数据矩阵 $\boldsymbol{Y}_1$,并将 $\boldsymbol{Y}_1$ 转换为 Stan 代码中定义的数据向量 $\boldsymbol{Y}$. 需要注意的是,所有在 Stan 程序数据模块中定义的变量都应该在列表 data_3pl 中明确赋值,否则 Stan 运行会报错. 最后一行的代码是使用"rstan"包调用 Stan 文件和已经定义的数据列表进行 MCMC 模型拟合抽样; 其中定义了 1 条马尔科夫链,迭代次数为 5 000; 类似于 BUGS 软件中的燃值(burn-in),Stan 中提供了一个热身值(warmup); "warmup = 2 500"意味着迭代次数前 2 500 次将会被舍去以获得更加精确的后验分布,默认值为迭代次数的 1/2; 函数"cores"等于 4 表示使用计算机的 4 个核心同时进行运算以提高运行速

度,可根据电脑配置的实际情况自主设定.

Stan 程序运行的结果将会保存在一个 S4 类的对象(object of S4 class)中,其中包括所有模型参数的后验分布信息. 例如,3PL 模型拟合的所有结果保存在 fit_3pl 的对象中,可以通过函数 traceplot 画抽样轨迹图检查模型的收敛性; 其中 inc_warmup 用于确认 是否在轨迹图中包括热身值的抽样过程(见图 1).

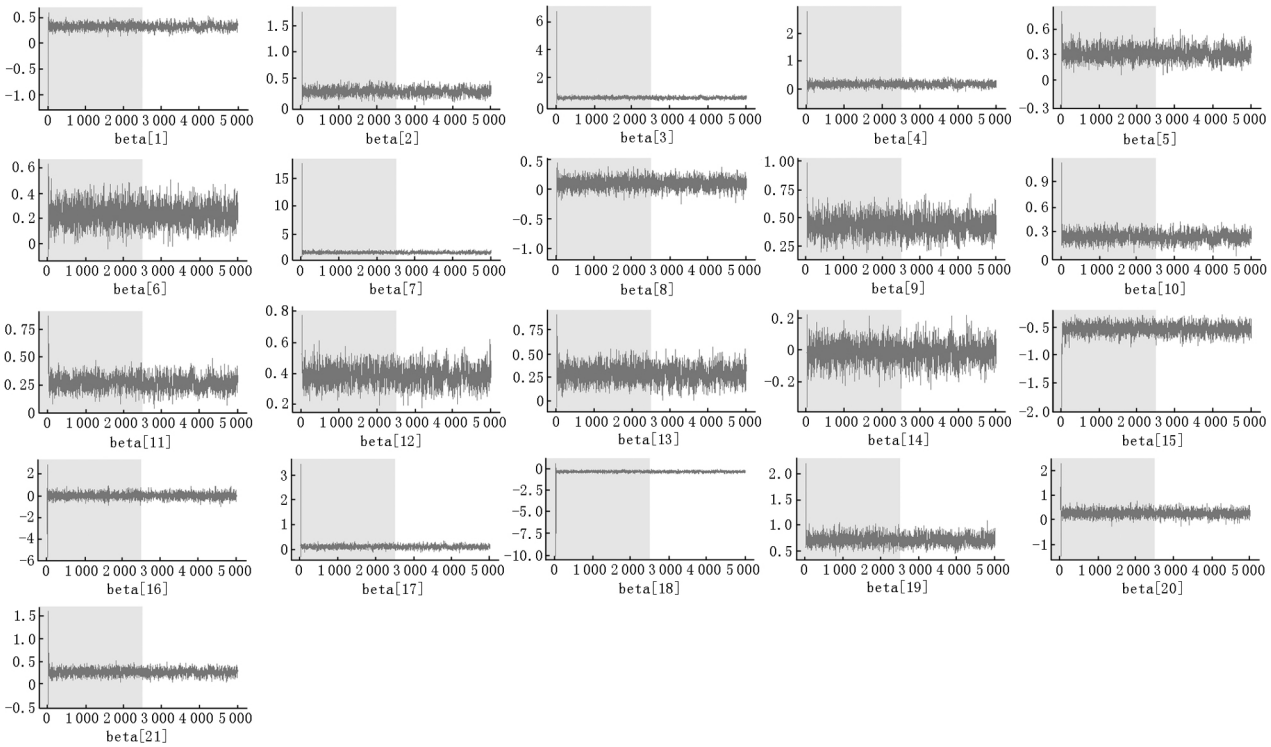

图 1 在心理类型量表数据中 2 参数 Logistic 模型(2PL)项目阈值参数的轨迹图

也可以使用函数 print 直接在 R 控制台打印 fit_3pl 中的有关所有参数的后验分布的摘要信息,值得注意的是表 1 中打印输出的结果与 WinBUGS 和 OpenBUGS 等传统贝叶斯统计软件输出的结果类似,均提供了估计参数的后验均值(mean)、标准差(standard deviation)等; n_eff 表示的是模型参数的后验分布抽样的有效样本量; Rhat 是由 S. P. Brooks 等[29] 改进的 Gelman-Rubin 收敛统计量,用以评价各参数的收敛情况. 当模型参数的后验分布的 $\hat{R}$ < 1.1 时,模型参数收敛,代码为

```
traceplot(fit_3pl,par = "alpha",inc_warmup = FALSE)
traceplot(fit_3pl,par = "beta",inc_warmup = FALSE)
traceplot(fit_3pl,par = "gamma",inc_warmup = FALSE)
print(fit_3pl,par = "theta")
print(fit_3pl,par = "alpha")
print(fit_3pl,par = "beta")
print(fit_3pl,pars = "gamma").
```

表 1 在心理类型量表数据中 2 参数 Logistics 模型(2PL)项目区分度参数的后验估计信息

alpha	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha[1]	1.54	0	0.16	1.25	1.43	1.53	1.64	1.88	1 903	1
alpha[2]	1.03	0	0.10	0.84	0.96	1.03	1.09	1.24	2 492	1
alpha[3]	0.97	0	0.11	0.77	0.90	0.97	1.04	1.20	1 794	1
alpha[4]	0.72	0	0.11	0.55	0.65	0.71	0.78	0.97	1 560	1
alpha[5]	0.85	0	0.09	0.69	0.79	0.85	0.91	1.03	2 389	1
alpha[6]	0.77	0	0.10	0.61	0.71	0.77	0.83	0.99	2 643	1
alpha[7]	0.61	0	0.08	0.47	0.56	0.61	0.66	0.78	1 890	1
alpha[8]	0.81	0	0.16	0.56	0.70	0.79	0.90	1.17	1 781	1
alpha[9]	0.81	0	0.09	0.64	0.75	0.81	0.87	1.02	2 519	1
alpha[10]	1.30	0	0.13	1.07	1.21	1.29	1.38	1.59	1 710	1
alpha[11]	1.30	0	0.13	1.05	1.21	1.29	1.38	1.57	1 980	1

表 1(续)

alpha	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha [12]	1.39	0	0.16	1.11	1.28	1.38	1.49	1.74	1 711	1
alpha [13]	0.76	0	0.11	0.58	0.68	0.75	0.82	0.99	2 281	1
alpha [14]	1.13	0	0.14	0.89	1.03	1.11	1.21	1.45	1 546	1
alpha [15]	0.85	0	0.12	0.65	0.76	0.83	0.93	1.13	1 581	1
alpha [16]	0.27	0	0.10	0.12	0.20	0.25	0.34	0.50	1 209	1
alpha [17]	1.09	0	0.12	0.89	1.01	1.08	1.17	1.35	2 079	1
alpha [18]	0.92	0	0.15	0.69	0.82	0.90	1.01	1.26	1 536	1
alpha [19]	0.75	0	0.08	0.60	0.70	0.74	0.80	0.92	2 440	1
alpha [20]	0.45	0	0.10	0.31	0.38	0.44	0.51	0.68	1 388	1
alpha [21]	0.97	0	0.10	0.79	0.90	0.96	1.03	1.17	2 347	1

确认模型参数收敛后,每个竞争模型的 WAIC 和 LOO 拟合指标均可以使用 R 包“loo”计算得到. 在 Stan 代码的预测值模块(generated quantity block)中已经计算出模型的对数似然(log-likelihood)并保存在 S4 类的对象(object of S4 class)中,可以使用函数 extract_log_lik 直接提取模型的后验分布对数似然,然后分别使用函数 loo 和 waic 计算出 LOO 和 WAIC 指标,代码为

```
log_lik <- extract_log_lik(fit_3pl)
loo_3pl <- loo(log_lik)
waic_3pl <- waic(log_lik).
```

表 2 列出了 3 个 IRT 竞争模型在真实数据中的拟合表现,从表 2 可看出 2PL 的 2 个拟合指标相比 1PL 和 3PL 2 个模型数值较小,这说明在 3 个 IRT 竞争模型中 2PL 与心理类型量表(MBTI)中的外向-内向维度数据拟合情况最为理想. 因此,在实际研究情境中,可以考虑采用 2PL 对真实数据的这个维度进行进一步分析.

表 2　3 种 2 级计分项目反应理论模型的 WAIC 和 LOO 指标

指标	1PL	2PL	3PL
LOO	18 576.3	18 084.9	18 117.7
WAIC	18 571.1	18 071.4	18 105.3

2.2　多维真实数据分析

多维真实数据研究采用的数据是高级瑞文推理智力测验第 II 套,因为 P. A. Carpenter 等[30] 只对测验共 36 个项目中的 25 个项目的规则进行了分类,故采用参数估计真实数据的项目数量为 25,被试人数为 886; 具体数据的详细信息读者可参见文献[15]. 项目测量的维度信息如图 2 所示.

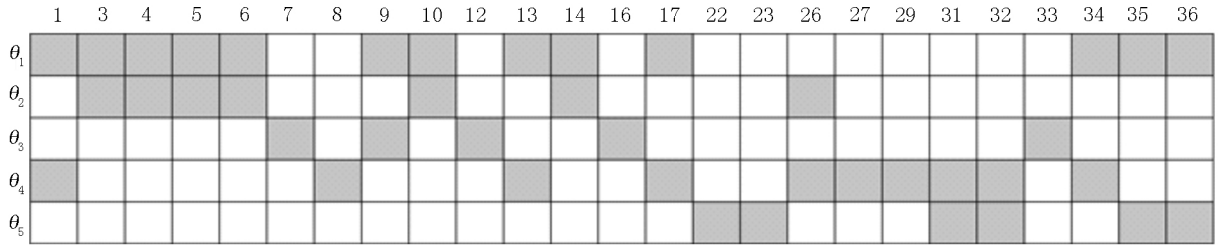

图 2　瑞文高级推理测验的维度测量 Q 矩阵

本节将采用 2 个多维 IRT 领域应用较为广泛的模型(多维 2 参数 Logistics 模型(M2PL)和多维 3 参数 Logistics 模型(M3PL))对高级瑞文推理智力测验真实数据进行竞争模型分析,M2PL 与 M3PL 可以通过增加或删除项目猜测参数相互转换,因此本文主要介绍相对复杂的模型 M3PL.

M. D. Reckase 等[31] 在回顾以往模型的基础上,最终提出了目前最实用的线性 Logistic 多维项目反应模型(M3PL). 该模型的表达式为

$$P(U_{ij}=1 \mid \boldsymbol{a}_j, c_j, d_j, \boldsymbol{\theta}_i) = c_j + (1-c_j)/(1+\exp(-1.7(\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i + d_j))), \quad (1)$$

其中 $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \cdots, \theta_{ik})^{\mathrm{T}}$ 为被试 i 的 k 维能力向量; $\boldsymbol{a}_j = (a_{j1}, a_{j2}, \cdots, a_{jk})^{\mathrm{T}}$ 为项目 j 的 k 维区分度向量; $\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i = a_{j1}\theta_{i1} + a_{j2}\theta_{i2} + \cdots + a_{jk}\theta_{ik}$; c_j 为项目的猜测度参数; d_j 为 MIRT 难度相关的参数,与单维 IRT 的难度参数 b_j 不同,但 2 者存在转换函数式

$$M_{DISC_j} = \sqrt{\sum_{k=1}^{K}(a_{jk})^2}, b_j = -d_j/M_{DISC_j},$$

M_{DISC_j} 反映项目 j 在多维能力空间上的整体区分度,它的值越大说明该项目在多维能力空间上的整体区分度越大,但并不一定代表它在每个能力维度上都

有较高的区分度. 该函数表达式里的 b_j 为 MIRT 的项目难度参数,它与 UIRT 的难度参数是同一概念,其值越大说明题目越难,反之越容易. 因此,对于测验为 k 维的 MIRT 模型来说,每个被试有 k 维能力,每个项目有 k 维区分度,但每个项目只有一个猜测度参数 c_j 和一个与项目难度相关的参数 d_j.

如前文所述,在默认 R 和 Stan 程序正确安装、运行环境正确配置的情况下,多维 IRT 模型的 Stan 代码已经被另外保存为".stan"后缀的独立文件,并和 R 文件、数据文件与项目测量的 Q 矩阵文件一同保存在 R 的默认工作目录下(用户可根据需要自行设定工作目录),数据文件保存的格式为不含缺失和异常值的数据矩阵. 与单维的 IRT 模型 Stan 代码类似,多维 IRT 模型的 Stan 代码也是分模块编写,数据(Data)模块中定义了真实数据的一些必要信息,如被试数量 K、观察作答数量 N、数据维度 D、观察作答的被试编号 jj、观察作答的项目编号 kk、数据矩阵 y、项目测量 Q 矩阵(Parameters)模块定义了模型(1)中的各个模型参数. 值得注意的是,因为模型(1)中的被试参数 θ 与项目区分度参数 a 的数据类型为矩阵,因此定义的函数与单维 IRT 模型代码中定义的方式不同,被试参数 θ 定义矩阵的函数为 matrix〈lower = -3,upper = 3〉[J,D] theta;θ 为取值在 [-3,3] 内的 J 行 D 列的矩阵. 项目参数 a 的定义矩阵函数为matrix〈lower = -0,upper = 3〉[D,K] alpha_ j; 同理 a 为取值范围在 [0,3] 内的D 行 K 列矩阵. 参数转换(Transformed parameters)模块定义了一个新的参数 alpha,根据模型测量的 Q 矩阵对已经定义的模型参数 alpha_ j 进行转换,之后的模型的定义与似然的计算均基于转换后得到新参数 alpha. 模型(Model)模块类似于单维 IRT 模型模块,主要定义 2 个部分: 先验(the prior)与似然(the likelihood);参数先验的定义与单维相同,即给出每个模型参数的先验分布,每个参数的先验分布可以参阅已有相关文献. 以下 2 行代码"to_vector(theta) ~ normal(0,1);to_vector(alpha_j) ~ lognormal(0,0.5);"表示组成矩阵 theta 的所有列向量均服从均值为 0、方差为 1 的标准正态分布,同理 alpha_ j 服从的是对数正态分布. 然后是定义模型,将模型(1)的公式定义在本模块中. 需要特别说明的是,函数 inv_logit(X) 等同于数学表达式"1/(1 + exp(-X));1.7* (theta [jj[i]] * col(alpha,kk[i])) + beta [kk [i]])"的数学表达式为 $1.7(a_j^{\mathrm{T}}\theta_i + d_j)$;最后,通过定义作答数据的概率分布定义似然函数. 预测值(Generated Quantities)模块中首先定义了 2 个新的向量类型的预测变量 vector [N] log_lik 和 vector [N] p,它们用于保存所有作答数据的对数似然和答对概率,因此数量均为 N 个;代码"bernoulli_lpmf(y [i] | p [i]);"表示的是作答 y 服从答对概率为 p 的伯努利概率质量函数(pmf)取对数.

本例中分别使用 2 个 2 级计分的多维 IRT 模型: M2PL 和 M3PL 对高级瑞文推理智力测验的真实数据进行分析,从表 3 中可以看出 M3PL 的 2 个拟合指标相对较小,这表明 M3PL 与真实数据的拟合情况较好. 因此,使用 M3PL 模型对真实数据进一步分析,估计真实数据的项目参数的后验分布信息; Stan 对于结果的反馈有多重呈现方式,R 包"shinystan"[13] 结合 Stan 能够提供更多的图形绘制选择,感兴趣的读者可以参照 M. K. Cowles 等[32] 的 MCMC 收敛诊断综述和 S. Sinharay[33] 的 IRT 应用研究. 图 3 与表 4 为 M3PL 拟合高级瑞文推理智力测验的部分输出结果,从图 3 和表 4 中可以看到 M3PL 中的难度相关参数 d 的拟合较好,所有项目的 Rhat 指标均为 1,这表明在设定条件下所有项目的 d 参数均收敛. 本例中 2 个多维模型的 R 代码除增加了定义模型维度 D 与项目测量 Q 矩阵外,其他设定与单维 3PL 模型相同. M3PL 的 R 代码为

```
Q <- read.csv("Data/Q_matrix.csv",header = F)
Data <- as.matrix(read.csv("Data/GJRW.csv"))
Y <- as.vector(Data)
J <- nrow(Data)
K <- ncol(Data)
D <- 5
N <- J* K
jj <- rep(1,K) %x% c(1:J)
kk <- c(1:K) %x% rep(1,J)
data_m3pl <- list(J = J,K = K,N = N,jj = jj,kk = kk,Y = Y,D = D,Q = Q)
fit_m3pl <- stan("M3PL.stan",data = data_m3pl,iter = 5 000,warmup = 2 500,cores = 4).
```

表 3 2 种 2 级计分的多维项目反应理论模型的 WAIC 和 LOO 指标

指标	M2PL	M3PL
LOO	18 136.3	18 082.7
WAIC	17 948.8	17 890.0

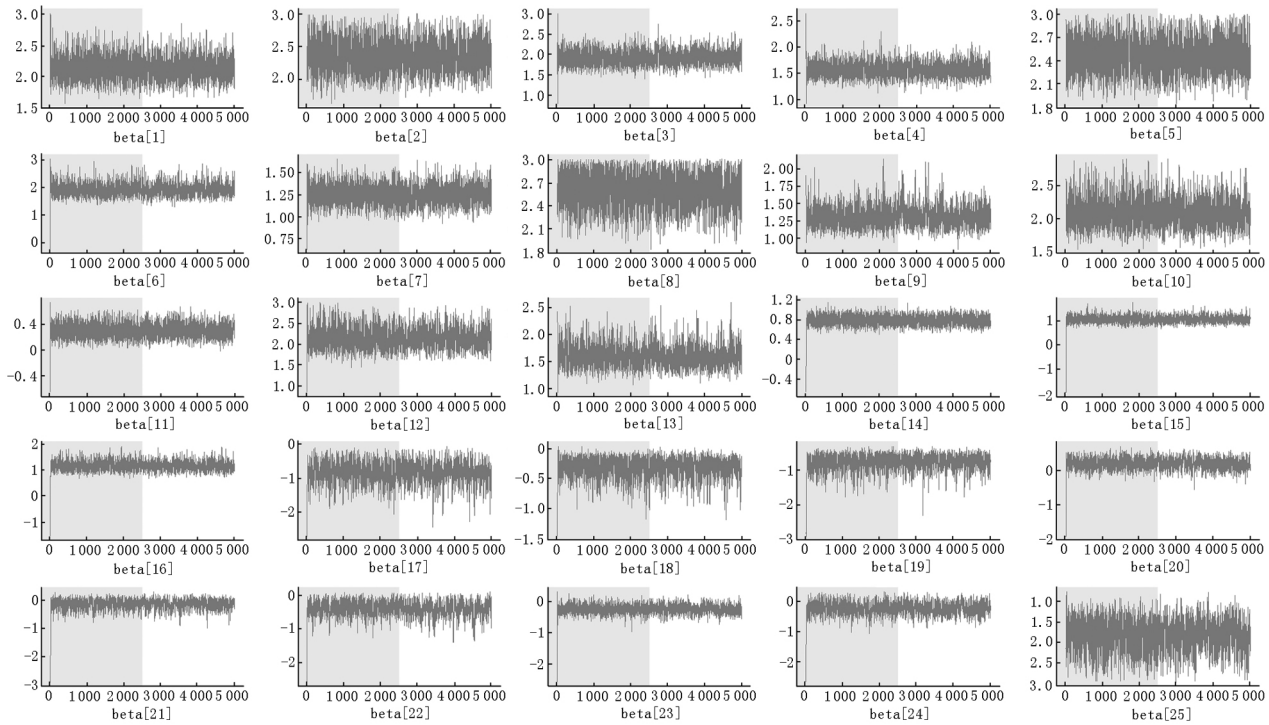

图 3　在高级瑞文推理智力测验数据中多维 3 参数 logistic 模型(M3Pl) 项目难度相关参数的轨迹图

表 4　在高级瑞文推理智力测验数据中 3 参数 logistics 模型(M3PL) 项目难度相关参数的后验估计信息

beta	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta [1]	2.47	0.01	0.24	2.00	2.29	2.46	2.64	2.95	1 286	1.00
beta [2]	2.54	0.01	0.23	2.08	2.37	2.54	2.72	2.96	1 904	1.00
beta [3]	1.84	0.00	0.15	1.58	1.74	1.83	1.93	2.15	2 790	1.00
beta [4]	1.75	0.00	0.16	1.46	1.64	1.74	1.85	2.11	1 898	1.00
beta [5]	2.55	0.00	0.20	2.17	2.40	2.55	2.69	2.94	2 240	1.00
beta [6]	1.71	0.01	0.18	1.41	1.59	1.69	1.82	2.13	854	1.00
beta [7]	1.27	0.00	0.11	1.07	1.19	1.27	1.34	1.49	2 978	1.00
beta [8]	2.44	0.01	0.24	2.00	2.26	2.42	2.61	2.92	1 632	1.00
beta [9]	1.20	0.00	0.11	1.00	1.12	1.19	1.26	1.42	3 000	1.00
beta [10]	2.15	0.01	0.24	1.72	1.98	2.13	2.31	2.69	1 366	1.00
beta [11]	0.26	0.00	0.13	0.00	0.16	0.25	0.35	0.52	1 123	1.00
beta [12]	1.86	0.01	0.20	1.53	1.73	1.84	1.97	2.32	988	1.00
beta [13]	1.48	0.02	0.24	1.15	1.33	1.44	1.57	1.97	202	1.01
beta [14]	0.79	0.00	0.09	0.61	0.73	0.79	0.86	0.98	2 626	1.00
beta [15]	1.35	0.01	0.22	0.96	1.20	1.34	1.49	1.81	1 078	1.00
beta [16]	0.96	0.00	0.13	0.72	0.87	0.96	1.05	1.22	1 358	1.00
beta [17]	−0.35	0.01	0.24	−0.89	−0.49	−0.31	−0.16	−0.01	286	1.00
beta [18]	−0.57	0.01	0.24	−1.11	−0.71	−0.54	−0.39	−0.20	553	1.00
beta [19]	−1.22	0.01	0.32	−1.94	−1.40	−1.18	−0.99	−0.70	735	1.00
beta [20]	0.15	0.01	0.15	−0.14	0.04	0.16	0.26	0.44	708	1.00
beta [21]	−0.07	0.00	0.11	−0.31	−0.13	−0.06	0.00	0.11	1 132	1.00
beta [22]	−0.31	0.00	0.16	−0.60	−0.43	−0.32	−0.20	−0.03	1 099	1.00
beta [23]	−0.12	0.01	0.17	−0.46	−0.24	−0.12	0.00	0.19	608	1.00
beta [24]	−0.16	0.00	0.15	−0.45	−0.26	−0.16	−0.07	0.13	963	1.00
beta [25]	−1.64	0.02	0.39	−2.55	−1.88	−1.61	−1.36	−0.96	555	1.00

3 讨论与总结

Stan 是一个相对较新的编程软件,它实现了相比较传统算法性能更加强大的 HMC 算法,与其他贝叶斯软件相比,Stan 更加高效,并且在估计一些复杂模型时所需的时间相比其他软件要少得多. 到目前为止,国内尚无系统介绍使用 Stan 进行贝叶斯项目反应理论模型参数估计的相关研究,因此对 Stan 在 IRT 领域中的应用方法进行介绍与分享是十分有必要的. 本文对使用 Stan 分析了项目反应理论领域应用较为广泛的 2 级计分的单维以及多维模型做了较为系统的介绍. 另外,本文也简要介绍了 2 个用于竞争模型比较的指标 WAIC 和 LOO,使用完全贝叶斯的方法,因此在理论上优于传统的 DIC 指标,并且使用 R 包"loo"能够较为容易地计算得到,对于统计基础较为薄弱的研究者来说更为友好,在 IRT 领域中应用并推广 Stan 不仅能丰富本领域的研究工具,为 IRT 模型的应用研究提供更为高效的技术支持,而且也为从事理论研究,特别是模型开发的研究者们提供了更加有效且友好的开发工具. 总的来说,Stan 在 IRT 领域中具有巨大的潜力,学习并推广 Stan 的应用技术对于国内 IRT 相关研究具有十分重要的积极意义.

当然,Stan 软件也存在一些限制与问题. 如对于缺失值的处理,传统的 WinBUGS 软件可以从缺失作答数据的后验预测分布中自动生成值,而 Stan 则需要另外在转换参数模块中定义它们,否则程序运行会报错误,整个过程是较为烦琐的. 当然,研究者也可以选择在使用 Stan 分析前,处理数据中的缺失值. 此外,虽然 Stan 使用较为便利,但其安装过程相比较传统 BUGS 软件较为烦琐,而且要求研究者对 R 语言等软件有一定的编程基础,与 BUGS 相比没有图形化的操作界面,因此在编写定义模型的过程中对于编程基础薄弱的研究者来说具有一定难度.

本文的初衷是向国内心理测量领域尤其是 IRT 和认知诊断领域的研究人员介绍使用 Stan 这一强大而高效的贝叶斯统计新工具进行贝叶斯项目反应分析的方法. 采用常见的 2 级计分的单维 IRT 和多维 IRT 模型进行真实数据的实例演示,希望能减少研究人员的 Stan 学习曲线,便于对进一步扩展的 IRT 模型进行模型参数估计. 对于那些希望使用 MCMC 算法将更复杂的 IRT 模型与数据进行拟合,但又不满足于 Gibbs sampler 和 Metropolis 算法在 BUGS 和其他 MCMC 包中的计算速度的研究人员来说,Stan 可能是一个相对更有吸引力的选择.

4 参考文献

[1] Carpenter B,Gelman A,Hoffman M D,et al. Stan: a probabilistic programming language [J]. Journal of Statistical Software,2017,76(1):1259-1270.

[2] Kruschke J. Doing Bayesian data analysis: a tutorial with R,JAGS,and Stan [M]. Pittsburgh: Academic Press,2014.

[3] Lunn D J,Thomas A,Best N,et al. WinBUGS a Bayesian modelling framework: concepts,structure,and extensibility [J]. Statistics and Computing,2000,10(4):325-337.

[4] Lunn D,Jackson C,Best N,et al. The BUGS book: a practical introduction to Bayesian analysis [M]. Chapman: Chapman and Hall/CRC,2012.

[5] Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling [EB/OL]. [2019-10-18]. http://core.ac.UR/display/20935350.

[6] Martin A D,Quinn K M,Park J H. MCMCpack: Markov chain monte carlo in R [EB/OL]. [2019-11-13]. https://www.oalib.com/paper/2884490#.XtoYuOSHodU.

[7] Geman S,Geman D. Stochastic relaxation,Gibbs distributions,and the Bayesian restoration of images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1984(6):721-741.

[8] Metropolis N,Rosenbluth A W,Rosenbluth M N,et al. Equation of state calculations by fast computing machines [J]. The Journal of Chemical Physics,1953,21(6):1087-1092.

[9] Neal R M. Probabilistic inference using Markov chain Monte Carlo methods [D]. Toronto,Ontario,Canada: University of Toronto,1993.

[10] Liang Zhongyao,Yu Yanhong,Wang Liqian,et al. A Bayesian ANOVA method to identify the temporal and seasonal dynamics of lake water quality variables [J]. Acta Scientiae Circumstantiae,2017,37(11):4170-4177.

[11] Neal R M. MCMC using Hamiltonian dynamics [J]. Handbook of Markov Chain Monte Carlo,2011,2(11):113-162.

[12] Banerjee S,Carlin B P,Gelfand A E. Hierarchical modeling and analysis for spatial data [M]. Chapmann: Chapman and Hall/CRC,2014.

[13] Stan Development Team. ShinyStan: interactive visual and numerical diagnostics and posterior analysis for Bayesian models [EB/OL]. [2019-11-13]. https://xueshu.baidu.com/usercenter/paper/show? paperid=652a67f6b2a3f9c616b0db8ab2d45280&site=xueshu_se.

[14] Stan Development Team. Stan modeling language users guide and reference manual [EB/OL]. [2019-11-13]. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=

4605C5A369842DE403C89BB7293BE5D8? doi = 10. 1. 1. 372. 3101&rep = rep1&type = pdf.

[15] 涂冬波,蔡艳,戴海琦,等. 多维项目反应理论: 参数估计及其在心理测验中的应用 [J]. 心理学报,2011,43(11):1329-1340.

[16] Curtis S M K. BUGS code for item response theory [J]. Journal of Statistical Software,2010,36(1):1-34.

[17] Stone C A,Zhu Xiaowen. Bayesian analysis of item response theory models using SAS [EB/OL]. [2019-11-13]. https://www.sas.com/content/dam/SAS/support/en/books/bayesian-analysis-of-item-response-theory-models-using-sas/67262_excerpt.pdf.

[18] 刘晋,汪秀琴,李天萍,等. 贝叶斯统计分析的新工具: Stan [J]. 中国卫生统计,2019,36(3):462-465,470.

[19] Rusch T,Maier M J,Hatzinger R. Linear logistic models with relaxed assumptions in R [EB/OL]. [2019-11-13]. https://link.springer.com/chapter/10.1007%2F978-3-319-00035-0_34.

[20] Team S D. RStan: the R interface to Stan [EB/OL]. [2019-11-13]. http://ftp.ps.pl/dsk0/CRAN/web/packages/rstan/vignettes/rstan_vignette.pdf.

[21] Swaminathan H,Gifford J A. Bayesian estimation in the three-parameter logistic model [J]. Psychometrika,1986,51(4):589-601.

[22] Vehtari A,Gelman A,Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC [J]. Statistics and Computing,2017,27(5):1413-1432.

[23] Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory [J]. Journal of Machine Learning Research,2010,11:3571-3594.

[24] Spiegelhalter D,Best N G,Carlin B P,et al. Bayesian measures of model complexity and fit [J]. Quality Control and Applied Statistics,2003,48(4):431-432.

[25] Akaike H. A new look at the statistical model identification [M]. New York: Springer,1974:215-222.

[26] Schwarz G. Estimating the dimension of a model [J]. The Annals of Statistics,1978,6(2):461-464.

[27] Luo Yong,Al Harbi K. Performances of LOO and WAIC as IRT model selection methods [J]. Psychological Test and Assessment Modeling,2017,59(2):183.

[28] 蔡华俭,朱臻雯,杨治良. 心理类型量表(MBTI)的修订初步 [J]. 应用心理学,2001,7(2):33-37.

[29] Brooks S P,Gelman A. General methods for monitoring convergence of iterative simulations [J]. Journal of Computational and Graphical Statistics,1998,7(4):434-455.

[30] Carpenter P A,Just M A,Shell P. What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test [J]. Psychological Review,1990,97(3):404.

[31] Reckase M D,McKinley R L. Some latent trait theory in a multidimensional latent space [EB/OL]. [2019-11-13]. https://files.eric.ed.gov/fulltext/ED264265.pdf.

[32] Cowles M K,Carlin B P. Markov chain Monte Carlo convergence diagnostics: a comparative review [J]. Journal of the American Statistical Association,1996,91(434):883-904.

[33] Sinharay S. Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples [J]. Journal of Educational and Behavioral Statistics,2004,29(4):461-488.

Using Stan to Implement Bayesian Parameter Estimation of IRT Models

LIU Siyang,CAI Yan*

(College of Psychology,Jiangxi Normal University,Nanchang Jiangxi 30022,China)

Abstract: Stan,a new probabilistic programming language for specifying statistical models,implements the powerful and efficient Hamiltonian Monte Carlo (HMC) sampling algorithm,which is significantly more efficient than the traditional Gibbs sampling and Metropolis algorithms. R package "rstan" links R and Stan,enabling Stan to run with R environment. First,this article introduces the programming language of Stan through the three-parameter logistic (3PL) model code. Secondly,this paper administrates Stan to calculate two new criteria of model-data fitting: WAIC and LOO,which provides an effective reference for IRT model studies. Finally,two types of real data are performed to investigate the performance of Stan in parameter estimation of one-dimensional IRT model and multidimensional IRT model respectively. In conclusion,this paper utilizes a new Bayesian statistical software to estimate the effectiveness and feasibility of this method through two empirical studies,which provides effective references for domestic scholars to apply Stan in IRT model research.

Key words: item response theory; HMC; Bayesian estimation; Stan　　(责任编辑: 冉小晓)