基于 GLM 的贝叶斯变量与模型选择①

汪建均，马义中
(南京理工大学经济管理学院，南京 210094)

摘要：针对非正态响应的部分因子试验，当筛选试验所涉及的因子数目较大时，提出了基于广义线性模型(generalized linear models，GLM)的贝叶斯变量与模型选择方法. 首先，针对模型参数的不确定性，选择了经验贝叶斯先验. 其次，在广义线性模型的线性预测器中对每个变量设置了二元变量指示器，并建立起变量指示器与模型指示器之间的转换关系. 然后，利用变量指示器与模型指示器的后验概率来识别显著性因子与选择最佳模型. 最后，以实际的工业案例说明此方法能够有效地识别非正态响应部分因子试验的显著性因子.

关键词：贝叶斯变量选择；部分因子试验设计；广义线性模型；筛选试验；非正态响应
中图分类号：F273.2 **文献标识码**：A **文章编号**：1007－9807(2012)08－0024－10

0 引 言

部分因子试验(fractional factorial experiments)能够有效地减少试验的次数，降低试验成本、缩短产品的研发周期，因此在工业试验设计中质量工程师倾向于选择部分因子试验设计. 通过科学的试验设计，筛选出显著性的可控因子，选择可控因子的最优水平搭配以减少或消除噪声因子对产品质量特性或工艺过程的影响，从而以较低的经济成本来提高和改善产品的质量[1]. 随着试验设计的广泛应用，研究者和质量工程师经常会遇到非正态响应，如泊松分布(产品的缺陷数)、二项分布(缺陷数比率)以及指数分布(如电子元器件的单位时间失效数)，因此非正态响应的部分因子试验设计在质量改进活动中显示出越来越重要的作用[2]. 特别是涉及到高质量、高可靠性的复杂产品，若无法精确地获得过程输入与输出之间的函数关系，将难以有效地筛选出合适的显著性因子. 因此，在实现复杂产品的质量设计中，应充分地利用试验数据的经验信息[3](如半导体制造过程中的电阻率满足伽玛分布等)，更为精确地刻画产品/过程输入与输出之间的函数关系，从而获得具有良好预测性能的响应模型.

1 国内外研究现状及分析

随着不断深入的研究，非正态响应的部分因子试验设计已引起了一些研究者的关注和重视. 传统方法是运用数据变换将非正态的数据转换为正态数据，然后运用最小二乘方法进行分析. 然而，一些研究者发现，难以寻找到合适的数据转换方法使得变换后的响应同时满足正态性、常数方差和系统线性可加 3 个基本条件[4]. 此外，基于数据转换方法的结果通常会造成预测精度较差，甚至出现与事实不相符合的情况[5]. 非正态响应的试验数据通常满足指数族分布，即方差是均值的函数关系. 广义线性模型(generalized linear models，GLM)不仅适用于广泛的指数族分布类型，而且能通过联系函数(link function)灵活地建立响应的方差与均值之间的函数关系. 针对非正态响

① 收稿日期：2010－07－05；修订日期：2011－08－19.
基金项目：国家自然科学基金重点资助项目(70931002).
作者简介：汪建均(1977—)，男，湖南张家界人，博士，讲师. Email: wjj19770818@163.com

应的试验，Lewis 等[6-7] 运用数据转换方法与 GLM 进行了对比分析，研究结果表明，GLM 能够得到更为合理的预测值以及更短的置信区间，因此 GLM 在参数估计和预测性能方面要优于数据转换方法. 尽管 GLM 在实现非正态响应的试验设计方面取得了一些积极的研究成果，但是在小样本、数据存在超散度(over-dispersion) 或者零膨胀(zero-inflation) 等情形下，GLM 的参数估计与变量筛选仍然存在一些问题没有很好解决[4] . Lewis 等[8] 运用 GLM 分析了仿真试验数据，研究结果表明，在全因子试验下，GLM 模型通常能够有效地识别所有显著性因子，但在部分因子试验时可能会出现无法有效识别某些显著性因子，甚至错误识别的情况.

近年来，由于贝叶斯方法在处理小样本数据以及利用先验信息方面的优势，国内外的一些学者尝试性地将其应用到非正态响应的工业试验设计中，并取得了一些研究成果. Weaver 和 Hamada[9] 以二项分布的试验数据为例，展示了如何利用贝叶斯方法进行模型选择、优化以及预测分析. Robinson 等[10] 运用贝叶斯方法分析了非正态响应的裂区试验(split-plot experiments) . 汪建均等[11] 运用 MCMC(Markov Chain Monte Carlo，MCMC) 方法对 GLM 进行了贝叶斯分析，并提出了两阶段的贝叶斯模型选择方法. 尽管 GLM 的贝叶斯分析能有效提高参数估计的可靠性与变量筛选的有效性，但以往的研究工作还是存在一些不足之处. 例如，未能全面地考虑模型的复杂度与预测性能等. 尤其是当筛选试验所涉及的因子数目较多时，在试验分析过程中往往还需要考虑因子之间的交互效应，因此需要考虑的模型个数将会相当大. 在这种情形下，试图拟合所有可能的模型，然后根据贝叶斯评估准则来筛选显著性变量，这样的操作方法在实际应用中难以有效地展开. 因此，必须寻求一种有效的算法来搜索整个模型空间(即所有候选变量不同组合所构成的模型) ，并计算这些候选变量与模型的后验概率，然后根据其大小来识别显著性因子或选择最佳模型. 针对一般线性回归模型，George 和 McCulloch[12] 提出了随机搜索的变量选择方法(stochastic search variable selection，SSVS) 来识别显著性变量与选择最佳模型. 针对变结构的线性回归模型，代逸生[13] 以贝叶斯准则为基础研究了在随机项方差相等和不相等两种情形下的贝叶斯模型选择问题. 针对小样本的试验数据和复杂的产品过程，崔庆安等[14] 提出基于支持向量机的响应曲面拟合方法，提高了模型的泛化能力与预测性能. Chipman 等[15] 运用 SSVS 方法分析了具有复杂混杂效应的试验，有效地识别显著性因子与选择最佳模型. Dellaportas 等[16] 在一定的情形下(如试验数据之间表现为正交关系) 假设模型参数之间条件独立，从而在 SSVS 方法的基础上提出了更为简洁的吉布斯变量选择 (Gibbs variable selection，GVS) 方法. Ntzoufras[17-18] 利用 WinBUGS 软件展示了 GVS 方法的有效性与优越性. 针对非正态响应的部分因子试验，本文在借鉴以往研究成果基础上提出基于 GLM 的贝叶斯变量与模型选择方法.

2　基于 GLM 的贝叶斯变量与模型选择

在广义线性模型中，无论是连续性响应变量还是离散性响应变量 y 均能表示为指数族(exponential family) 的概率分布形式，其基本表达式如下

$$f(\boldsymbol{y}) = \exp\{a^{-1}(\varphi)(\boldsymbol{y}\boldsymbol{\theta} - b(\boldsymbol{\theta})) + c(\boldsymbol{y},\varphi)\} \tag{1}$$

式中: $\boldsymbol{\theta}$ 为规范参数(canonical parameter) ; φ 为散度参数 (dispersion parameter) . 根据式(1) 中 $a(\varphi)$ 、$b(\boldsymbol{\theta})$ 和 $c(\boldsymbol{y},\varphi)$ 的不同形式能够构建出不同分布的函数. 有关 GLM 的基本理论与参数估计等具体问题可参考文献[5] .

在广义线性模型中，不同分布函数的模型均能通过相应的联系函数加以构建，因此 GLM 的建模关键在于如何选择合适的分布函数与联系函数. 分布函数的选择往往涉及到具体的试验数据特征，通常根据相关试验的信息或试验者的经验进行选择. 联系函数的选择相对更为灵活，可以根据需要选择不同形式的联系函数. 在这些联系函数中存在比较特殊的联系函数，即规范联系函数

(canonical link function). 它通过设置规范参数 θ 等于线性预报器 η,从而满足 $\theta = X\beta$. 由于通过规范联系函数所构造的联合密度函数相对其他联系函数所得到的结果更为简洁,因此选择规范联系函数对 GLM 进行统计推断相对更容易些.

在 GLM 框架下,变量选择的主要问题表现为线性预报器中的变量选择. 针对某个广义线性模型,其线性预报器 η 可以表示为

$$\eta_i = \sum_{j=0}^{p} \gamma_j x_{ij}\beta_j \tag{2}$$

式中: γ_j 为二元变量指示器. 如果二元变量指示器 γ_j 等于 1,则相应的变量 x_j 包含在模型中; 如果二元变量指示器 γ_j 等于 0,则相应的变量 x_j 将从模型中被剔除出去. x_{ij} 为向量 X_i 的第 j 个变量,β_j 为第 j 个变量所对应的估计参数.

2.1 考虑经验贝叶斯先验的变量选择方法

为了在 GLM 框架下有效实施贝叶斯变量选择方法,针对参数 $\theta = (\gamma,\beta)$ 选择采用分层结构形式 $f(\gamma,\beta) = f(\gamma)f(\beta \mid \gamma)$ 来设定先验. Ntzoufras [18] 提出可根据变量是否包含在模型中将参数 β 划分为 $\beta_\gamma(\gamma_j = 1)$ 和 $\beta_{\setminus\gamma}(\gamma_j = 0)$,从而更方便地利用 MCMC 方法对 GLM 进行贝叶斯分析. 根据 GVS 方法设定参数先验的经验,将先验分布 $f(\beta \mid \gamma)$ 划分真实参数先验 $f(\beta_\gamma \mid \gamma)$ 与伪先验(pseudo-prior) $f(\beta_{\setminus\gamma} \mid \beta_\gamma,\gamma)$. 针对所有变量指示器 γ_j,在缺乏相关先验信息时通常假定为满足相互独立且概率为 0.5 的 Bernoulli 分布. 在上述假定条件下,模型参数的满条件后验(full conditional posterior) 分布为

$$f(\beta_\gamma \mid \beta_{\setminus\gamma},\gamma,y) \propto f(y \mid \beta_{\setminus\gamma},\beta_\gamma,\gamma) \times f(\beta_{\setminus\gamma} \mid \beta_\gamma,\gamma) f(\beta_\gamma \mid \gamma) \tag{3}$$

$$f(\beta_{\setminus\gamma} \mid \beta_\gamma,\gamma,y) \propto f(\beta_{\setminus\gamma} \mid \beta_\gamma,\gamma) \tag{4}$$

$$f(\gamma_j \mid \beta_{\setminus\gamma},\beta_\gamma,y) \sim \text{Bernoulli}(0.5),\ j = 1,\cdots,p \tag{5}$$

从式(3) 可知,参数 β_γ 的满条件后验分布依赖于 $f(\beta_{\setminus\gamma} \mid \beta_\gamma,\gamma)$. 若变量之间存在较强的相关性时考虑上述依赖关系是有意义,但通常假定模型参数 β_γ 与 $\beta_{\setminus\gamma}$ 之间是相互独立以简化条件后验分布. 则式(3) 与式(4) 可以进一步得到简化,其具体形式如下

$$f(\beta_\gamma \mid \beta_{\setminus\gamma},\gamma,y) \propto f(y \mid \beta_{\setminus\gamma},\beta_\gamma,\gamma) \times f(\beta_{\setminus\gamma} \mid \gamma) f(\beta_\gamma \mid \gamma) \tag{6}$$

$$f(\beta_{\setminus\gamma} \mid \beta_\gamma,\gamma,y) \propto f(\beta_{\setminus\gamma} \mid \gamma) \tag{7}$$

在 GLM 模型的贝叶斯分析中,参数的先验分布选择不恰当可能会造成不正常的后验分布,因此合理选择参数的先验分布就显得非常重要. 特别是在考虑诸多分布的 GLM 框架下,对估计参数选择合适的先验分布需要更加谨慎. Ibrahim 和 Laud [19] 给出在 GLM 模型中运用 Jeffreys 先验信息的充分必要条件,并指出在 GLM 框架下运用 Jeffreys 先验信息,可能会比其他无信息先验得到更为可靠和精确的结果. 笔者利用 Jeffreys 先验对 GLM 进行了贝叶斯分析与参数估计,研究发现,在所有参数都满足收敛性诊断的情况下,参数 β 的后验分布核密度曲线非常光滑,且近似服从正态分布 [11]. 在这种情形下,针对 GLM 的参数 β 选择经验贝叶斯混合正态先验,其具体形式如下

$$f(\beta_j \mid \gamma_j,y) = \gamma_j N(\mu_{\beta_j},S^2_{\beta_j}) + (1-\gamma_j) N(\bar{\mu}_{\beta_j},\bar{S}^2_{\beta_j}) \tag{8}$$

式中: μ_{β_j} 和 $S^2_{\beta_j}$ 是真实参数 β_γ 所对应的先验均值与先验方差; $\bar{\mu}_{\beta_j}$ 和 $\bar{S}^2_{\beta_j}$ 是伪参数 $\beta_{\setminus\gamma}$ 所对应的先验均值与先验方差. 在此,首先利用 Jeffreys 先验对全模型(包含所有变量的 GLM) 进行贝叶斯分析与参数估计,然后根据各参数的后验核密度曲线以判断各参数是否满足正态分布的假定. 若 GLM 的各参数满足收敛性诊断与正态先验设定,则可以假设伪参数 $\beta_{\setminus\gamma}$ 的先验均值 $\bar{\mu}_{\beta_j}$ 和方差 $\bar{S}^2_{\beta_j}$ 等于由 SAS GENMOD 程序 [20] 拟合全模型所得到的后验均值 $\hat{\mu}_{\beta_j}$ 与后验方差 $\hat{S}^2_{\beta_j}$. 借鉴 Ntzoufras 所提出的经验贝叶斯先验设定方法 [18],可以假定真实参数 β_γ 的先验均值 μ_{β_j} 等于后验均值 $\hat{\mu}_{\beta_j}$,而先验方差 $S^2_{\beta_j}$ 等于 $N\hat{S}^2_{\beta_j}$,其中 N 为试验设计的样本量大小. 当样本量较小时,根据试验数据所能够获得的样本信息相对较弱,而模型参数的先验信息对参

数后验的影响会相对更大些,因此在先验方差中考虑样本量 N 的大小,可以有效地调节样本量大小对试验结果的影响. 根据以上经验贝叶斯先验分布,各参数的均值与方差的先验可进一步简化为

$$\beta_j \sim N(\mu_{\beta_j,\gamma}, S^2_{\beta_j,\gamma}) \tag{9}$$

$$\mu_{\beta_j,\gamma} = \hat{\mu}_{\beta_j},\quad S^2_{\beta_j,\gamma} = \{\gamma_j N + (1-\gamma_j)\}\hat{S}^2_{\beta_j} \tag{10}$$

以上的经验贝叶斯先验方法不仅考虑了样本量对试验结果的影响,而且充分利用了贝叶斯方法能进行信息更新的优势,有效地将历史数据信息转化为各参数的先验信息.

2.2　基于贝叶斯后验样本的统计推断

根据以上各参数的先验设定方法,通过 MCMC 方法动态模拟各参数后验分布的马尔科夫链,并根据各参数的后验样本进行收敛性诊断与参数估计. 然后,贝叶斯变量选择方法的主要工作是如何有效地估计各变量指示器与所有可能模型的后验概率. 针对每个变量指示器,建立二元变量指示器 γ 与十进制模型指示器 $M(\gamma)$ 的转换关系,其基本形式如下

$$M(\gamma) = \sum_{j=k}^{p} \gamma_j 2^{j-k} \tag{11}$$

若常数项包含在所有可能的模型之中,则 $k=1$; 否则 $k=0$. 根据 MCMC 方法动态模拟所得到的变量指示器 γ 与模型指示器 $M(\gamma)$ 的后验样本,可以估计得到每个变量指示器 γ 的后验概率 $f(\gamma=1\mid y)$ 与模型指示器的后验概率 $f(M\mid y)$,其计算公式如下

$$\hat{f}(\gamma=1\mid y) = \frac{1}{T-B_t}\sum_{t=T+1}^{T} I(\gamma^{(t)}=1) \tag{12}$$

$$\hat{f}(M\mid y) = \frac{1}{T-B_t}\sum_{t=T+1}^{T} I(M^{(t)}=M) \tag{13}$$

式中: T 和 B 分别是后验样本的全部迭代次数和燃烧期(burn-in) 间的迭代次数,$\gamma^{(t)}$ 和 $M^{(t)}$ 是变量指示器与模型指示器所对应的后验样本在第 t 次的迭代结果. 当筛选试验所考虑的因子(含交互效应) 数目较大时,需要考虑的模型数目将会非常大,则上述方法无法有效地计算所有可能模型的后验概率.

Barbieri 和 Berger[21] 指出,在某些情况下,中位数概率模型(变量指示器的后验概率大于或等于 0.5,即 $\hat{f}(\gamma=1\mid y)\geqslant 0.5$) 可能比最大后验模型有更好的预测性能. 尽管中位数概率模型在很多情形下能够有效地缩减模型空间,但是还是存在相当大的不确定性. 譬如,当某个变量指示器的后验概率稍微小于 0.5(如 $\hat{f}(\gamma_i=1\mid y)=0.49$) 时,根据中位数概率方法将认为该变量指示器所对应的变量为不显著性变量,在这种情形下所得到的中位数概率模型未必具有最佳的预测性能. 因此,中位数概率模型在某些情形下具有相当大的不确定性. Fouskkis 等[22] 提出可以根据变量指示器的后验概率大小来缩减所考虑的模型空间. 当变量指示器的后验概率较小 (如 $\hat{f}(\gamma=1\mid y)<0.3$) 时,可以剔除对应的变量以减少所考虑的模型个数. 然而,这种方法只能部分地缩减模型的空间,而且无法有效地得到最大后验概率的模型. 针对以往的研究不足,本文提出的根据变量指示器的后验概率来缩减模型空间的新方法,其具体实施步骤如下.

1) 考察先验设定的合理性. 利用 Jeffreys 先验对全模型(包含所有变量的 GLM) 进行贝叶斯分析与参数估计,然后根据各参数的后验核密度曲线判断各参数是否满足正态分布的假定[11].

2) 设定参数的经验贝叶斯先验. 针对各参数设定经验贝叶斯先验分布,并运用 MCMC 方法对 GLM 进行参数估计与收敛性诊断.

3) 计算变量指示器的后验概率. 根据各变量指示器的后验样本,利用式(12) 计算各变量指示器的后验概率.

4) 确定变量的显著性程度. 根据各变量指示器的后验概率大小区分变量的显著性程度. 若变量指示器的后验概率很高(如后验概率在 0.9 以上),则这些变量指示器所对应的变量将视为高度显著性变量,将包含在最终的模型之中; 若变量指示器的后验概率较低(如后验概率在 0.3 以下),则这些变量指示器所对应的变量将视为不显著性变量,将从最终的模型中剔除出去; 若变量指示器的后验概率较高(如后验概率在 0.3 ~ 0.9

之间)，则这些变量指示器所对应的变量将视为相对显著性变量. 上述经验数据的设定参考以往研究的结果，其主要目的是缩减需要考察的模型空间，以提高 WinBUGS 程序的运行速度与监控效果.

5) 计算候选模型的后验概率. 根据第 4) 步所得到的相关信息，利用贝叶斯信息更新的思想对变量指示器的先验概率进行重新设定. 考虑到高度显著性变量将始终包含在最终模型之中，因此可以假设其变量指示器的先验概率为 1.0; 不显著性变量将从最终模型中剔除出去，因此其变量指示器的先验概率可以假设为 0; 相对显著性变量还需要进一步考察，因此其变量指示器的先验概率可以假设为 0.5. 针对相对显著性变量，运用 WinBUGS 程序计算所有可能模型(若候选变量个数为 q，则所有可能的模型个数为 2^q) 的后验概率.

6) 确定显著性变量与最佳模型. 根据第 5) 步计算的结果，比较所有候选模型的后验概率，进而筛选出所有显著性变量，确定最佳模型.

3 实例分析

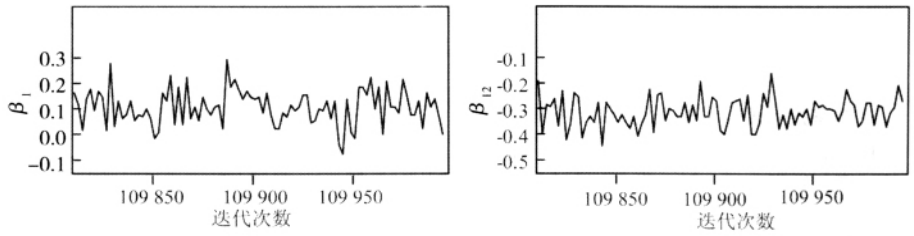
在波峰(wave soldering) 焊接过程中，将电子元件装配到线路板上时经常会出现焊接接口缺陷，这是电子元件制造过程中常见的质量问题. 整个焊接过程包括烘烤和预热电路板，然后用传送机将电路板通过焊接波峰. 为了减少焊接缺陷，试验者实施了 2_{IV}^{7-3} 的部分因子试验，并定义了混杂关系: $E = ABD$, $F = ACD$, $G = BCD$. 这些试验因子分别为: 预烤温度(A)，焊接密度(B)，传送机速度(C)，预热条件(D)，冷却时间(E)，超声波焊接搅拌器(F) 以及焊接温度(G). 响应 y 是某个电路板上的缺陷个数. 对每个因子水平组合，该试验评估了 3 个电路板上的缺陷. 为了考察新方法在小样本情形下识别显著性因子的效果，本文取每个水平组合下 3 个电路板上缺陷数的平均值作为模型的响应，具体的试验数据见参考文献[3](第 13 章，p564). 试验者根据以往的经验认为存在可能重要的 6 个交互效应，分别为 AB, AC, AD, BC, BD 和 CD. 根据上述的试验信息，首先拟合饱和的广义线性模型(响应为泊松分布)，包含 7 个主效应($A - G$)、6 个交互效应 AB, AC, AD, BC, BD 和 CD 以及两个附加效应 CE 和 ABC. 有关模型的参数估计以及收敛性的诊断结果可参考文献[11]. 尽管 GLM 的贝叶斯分析能有效地提高参数估计的稳健性与显著性因子筛选的可靠性，但是以往的研究仍有一些不足之处，如未能全面地考察整个模型的空间等. 针对以往研究的不足，本文运用所提的新方法进一步地加以改进.

考虑到所给实例的数据特征，根据式(5) 构建一个含变量指示器 γ 的对数联系函数，其具体形式为

$$\begin{aligned}\log(\mu) = {} & \gamma_0\beta_0 \times \mathrm{intercept} + \gamma_1\beta_1 \times A + \\ & \gamma_2\beta_2 \times B + \gamma_3\beta_3 \times C + \gamma_4\beta_4 \times D + \\ & \gamma_5\beta_5 \times E + \gamma_6\beta_6 \times F + \gamma_7\beta_7 \times G + \\ & \gamma_8\beta_8 \times AB + \gamma_9\beta_9 \times AC + \\ & \gamma_{10}\beta_{10} \times AD + \gamma_{11}\beta_{11} \times BC + \\ & \gamma_{12}\beta_{12} \times BD + \gamma_{13}\beta_{13} \times CD + \\ & \gamma_{14}\beta_{14} \times CE + \gamma_{15}\beta_{15} \times ABC \quad (14)\end{aligned}$$

其中: γ_j 为二元变量指示器; 系数 β_j 为模型中各变量所对应的估计系数. 根据文献[11] 的计算结果，可以得到所拟合饱和模型各参数的后验均值与方差 $\hat{S}^2_{\beta_j}$(即为后验标准差的平方). 针对新模型各参数设置经验贝叶斯混合正态先验，利用 WinBUGS 软件设计相应的程序实现 GLM 的参数估计与收敛性诊断. 首先，针对各参数进行 10 000 次的初始迭代，舍弃这些燃烧期的迭代样本以保证参数的收敛性诊断; 其次，为减少各参数迭代抽样值之间的自相关性，在 WinBUGS 程序中设置 thin = 2，即每间隔 2 步抽样 1 次; 然后，再进行 100 000 次的迭代以获得具有良好收敛性的样本，并利用这些样本进行收敛性诊断与参数估计. 考虑到试验结果的相似性，仅给出参数 A 与参数 BD 所对应系数 β_1 与 β_{12} 的抽样值踪迹图、核密度曲线图与自相关图，其结果分别如图 1、图 2 和图 3 所示. 根据各参数所对应系数的抽样值踪迹图、核密度曲线图以及自相关图，可以直观地判断出各参数均收敛于相应的稳态分布，且满足参数收敛性诊断要求. 从各参数的核密度曲线图可以看出其曲线非常光滑且具有高度的对称性，非常近似为

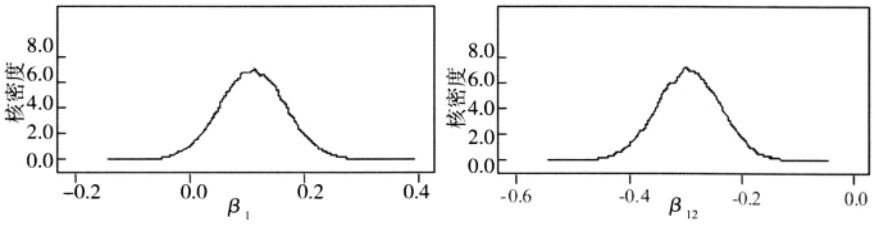
正态分布.



(a) β_1　　(b) β_{12}

图 1　参数 A 和参数 BD 所对应系数 β_1 和 β_{12} 迭代抽样值的踪迹图

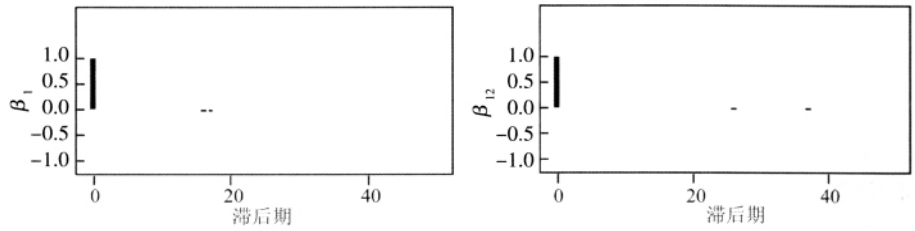
Fig. 1 Trace plots for β_1 and β_{12} corresponding to A and BD



(a) β_1　　(b) β_{12}

图 2　参数 A 和参数 BD 所对应系数 β_1 和 β_{12} 的后验密度曲线图

Fig. 2 Posterior density plots for β_1 and β_{12} corresponding to A and BD



(a) β_1　　(b) β_{12}

图 3　参数 A 和参数 BD 所对应系数 β_1 和 β_{12} 的迭代抽样值自相关图

Fig. 3 Autocorrelation plots for β_1 and β_{12} corresponding to A and BD

为了进一步考察各变量的显著性程度,在各参数满足收敛诊断后估计各变量指示器的后验概率,具体的计算结果如表 1 所示. 根据表 1 中各变量指示器的后验均值与前述的经验数据,将所有变量划分为 3 个子集,分别为高度显著的变量集,包含变量 B, C, G, AC, BD; 不显著的变量集,包含变量 D, F, AB, CD, CE, ABC. 相对显著的变量集,包含变量 A, E, AD, BC. 针对相对显著的候选变量,进一步利用 WinBUGS 软件设计相应的算法,计算所有候选模型的后验概率. 根据本文所提方法的第 5) 步,针对式(5) 重新设置如下的先验信息,其具体形式如下

$$f(\gamma_j \mid \beta_j, y) \sim \text{Bernoulli}(\text{pgamma}[j]),\quad j = 1,\cdots,15$$

```
pgamma = c(0.5,1,1,0,0.5,0,1,0,1,
0.5,0.5,1,0,0,0)
```

针对候选变量 A, E, AD, BC,利用式(11) 与式(13) 在原 WinBUGS 程序中加入相应的代码,其具体形式如下

```
model ← gamma[1] * gamma[5] * 2 +
        gamma[10] * 4 + gamma[11] * 8
for(m in 1 : 16){ pmodel[m] ←
        equals(model + 1, m)
```

其中 gamma[ ] 分别代表候选变量 A, E, AD, BC 所对应变量指示器的后验抽样值. WinBUGS 软件的内置函数 equals 可以判断候选模型与目标模型(按候选变量顺序排列好的模型) 是否一致,从而方便地统计所有候选模型的后验概率. 按后验概率大小依次排序,其结果如表 2 所示.

表 1 各变量指示器的后验统计量

Table 1 Posterior statistic summaries of each variable indicator

变量指示器	后验均值	模拟误差	后验分位数		
			2.5%	50%	97.5%
γ_0	**1.000 0**	**4.472E - 13**	**1.00**	**1.00**	**1.00**
γ_1	0.625 6	0.002 3	0.00	1.00	1.00
γ_2	**0.970 6**	**7.218E - 4**	**0.00**	**1.00**	**1.00**
γ_3	**1.000 0**	**4.472E - 13**	**1.00**	**1.00**	**1.00**
γ_4	0.197 5	0.001 8	0.00	0.00	1.00
γ_5	0.461 1	0.002 3	0.00	0.00	1.00
γ_6	0.202 9	0.001 7	0.00	0.00	1.00
γ_7	**1.000 0**	**4.472E - 13**	**1.00**	**1.00**	**1.00**
γ_8	0.264 2	0.002 0	0.00	0.00	1.00
γ_9	**0.983 5**	**5.813E - 4**	**1.00**	**1.00**	**1.00**
γ_{10}	0.756 8	0.001 9	0.00	1.00	1.00
γ_{11}	0.477 6	0.002 4	0.00	0.00	1.00
γ_{12}	**1.000 0**	**4.472E - 13**	**1.00**	**1.00**	**1.00**
γ_{13}	0.290 2	0.002 0	0.00	0.00	1.00
γ_{14}	0.202 2	0.001 9	0.00	0.00	1.00
γ_{15}	0.214 5	0.001 8	0.00	0.00	1.00

注:表中黑体部分为高度显著性变量(根据本文所提出的新方法判断)的统计结果.

根据表 2 的计算结果可知,包含常数项、高度显著性变量 B, C, G, AC, BD 以及候选变量 A, AD 的模型 1 具有最大的后验概率. 从表 1 的计算结果可知,最大后验概率模型所包含的变量都满足 $\hat{f}(\gamma = 1 \mid y) > 0.5$,即这些变量所对应的变量指示器后验概率均大于 0.5. 此时,根据表 1 所得到的中位数概率模型与表 2 所得到的最大后验概率模型是完全一致的. 若根据式(5)更新变量指示器的后验概率,重新运行 WinBUGS 程序可以得到变量 BC 的后验概率为 0.498 3. 从表 2 的计算结果也可以看出,包含常数项、高度显著性变量 B, C, G, AC, BD 以及候选变量 A, AD 和 BC 的模型 2 也具有较高的后验概率,与最大后验概率模型的结果非常接近. 二者之间的差异在于候选变量 BC 是否应该包含在最终模型之中. 无论是从表 2 的计算结果,还是从更新变量指示器先验所得到的结果,都可以明显地看出变量 BC 所对应的变量指示器后验概率非常接近 0.5. 综合以上的分析,本文倾向于选择表 2 中的模型 2 为最终模型,其对应的显著性变量分别为常数项,变量 A, B, C, G, AC, AD, BC 和 BD. 这一研究结果与参考文献[11]的研究结果完全一致,因此本文所提出的新方法能够有效地识别出非正态响应部分因子试验的显著性因子. 此外,本文所提出的方法还可以进一步考察其他候选变量是否应该包含在最终模型,进而提高模型的预测能力. 从表 1 与表 2 的计算结果可知,本文的研究结果很好地支持 Barbieri 和 Berger[21] 所提出的中位数概率模型. 当筛选试验所涉及的因子数目特别大,特别是根据变量指示器与经验数据所得到的候选变量数目仍然很大时,可以先通过中位数概率模型的方法进行简化处理,有效地筛选出显著性变量;然后针对某些变量指示器的后验概率接近 0.5 的变量进行重点分析,计算其相应模型的后验概率值,进而确定最佳模型. 需要特别指出的是,当候选变量较多时,前述的目标模型(按候选变量顺序排列好的模型)与候选变量之间的关系难以有效地一一对应. 为此,根据

二元变量指示器与十进制模型变量指示器的转换关系,专门设计相应的 R 程序以帮助输出目标模型以及相应的模型指示器数值,即表 2 中的第 2 列与第 3 列结果. 考虑到篇幅有限,本文未给出所提方法的实现程序,若需要可以通过电子邮件与第一作者联系获得.

表 2　候选模型的后验概率结果

Table 2 Posterior probability summaries of the candidate models

次序	模型	模型指示器	模型后验概率	模拟误差
1	***A + AD***	**6**	**0.181 0**	**0.005 9**
2	$A + AD + BC$	14	0.170 4	0.004 7
3	$A + E + AD + BC$	16	0.106 0	0.004 1
4	$A + E + AD$	8	0.091 8	0.003 9
5	$E + AD + BC$	15	0.084 0	0.004 1
6	$E + AD$	7	0.082 6	0.003 8
7	AD	5	0.072 2	0.003 2
8	$AD + BC$	13	0.068 0	0.003 4
9	$A + BC$	10	0.028 6	0.002 7
10	$A + E + BC$	12	0.026 8	0.002 2
11	A	2	0.020 2	0.002 0
12	$E + BC$	11	0.018 6	0.001 9
13	E	3	0.018 2	0.002 2
14	$A + E$	4	0.016 4	0.001 7
15	BC	9	0.008 4	0.001 3
16	仅含高度显著性变量与常数项	1	0.006 8	0.001 1

注:1. 表中所有模型均含有高度显著性变量与常数项; 2. 黑体部分为最大后验概率模型的统计结果.

4　结束语

在质量设计研究中,通常会考虑经济性与可操作性等因素的影响选择部分因子试验设计,但传统的建模方法(如数据转换、响应曲面模型等)在非正态部分因子试验的情形下往往难以有效识别出所有显著性因子. 针对非正态部分因子试验设计,特别是当筛选试验所涉及的因子数目较大时,提出基于 GLM 的贝叶斯变量与模型选择方法. 该方法有效地解决了筛选试验因子数目较大时变量筛选与模型选择的问题,扩展了非正态响应的变量选择方法与建模技术. 该方法在建模过程中对模型参数与变量指示器的先验进行了多次更新,最大限度地利用相关的试验数据信息. 此外,该方法还充分地借鉴了随机搜索算法的思想,全面地考察了所有可能模型的后验概率,为变量筛选与模型选择提供了科学的决策依据. 基于 GLM 的贝叶斯变量选择方法不仅能够有效地刻画各种常见指数族分布,而且能够利用贝叶斯方法的优势处理小样本的数据问题,因此本文所提出的方法不仅适用于质量设计领域,而且能广泛地应用于管理科学、工程科学等其他领域的变量筛选与模型构建.

本文的研究是建立对试验因子的混杂效应具有一定了解,对需要重点考察试验因子的交互效应具有一定先验知识基础上展开的. 若缺乏这些先验信息时,可以结合因子效应的 3 个基本原则[3](即效应排序原则,effect hierarchy principle; 效应稀疏原则,effect sparsity principle; 效应遗传原则,effect heredity principle)以获得试验因子之间的先验信息. 利用贝叶斯方法解决具有高度混杂效应的非正态响应质量设计问题,将是未来需要进一步研究的课题.

参 考 文 献:

[1] 马义中，赵逢禹. 多元质量特性的稳健设计及其实现 [J]. 系统工程与电子技术，2005，27(9)：1580－1582.
Ma Yizhong，Zhao Fengyu. Robust design and implementation for multivariate quality characteristics [J]. Systems Engineering and Electronics，2005，27(9)：1580－1582. (in Chinese)

[2] Myers R H，Montgomery D C. A tutorial on generalized linear models [J]. Journal of Quality Technology，1997，29(3)：274－291.

[3] Myers R H，Montgomery D C，Vinning G G. Generalized Linear Models with Application in Engineering and the Sciences [M]. New York：John Wiley & Sons，2002.

[4] Wu C F J，Hamada M. Experiments：Planning，Analysis，and Parameter Design Optimization [M]. New York：John Wiley & sons，2000.

[5] 何 桢，马彦辉，赵 有. 非正态响应的部分因子试验设计 [J]. 工业工程，2008，11(2)：72－75.
He Zhen，Ma Yanhui，Zhao You. On fractional factorial experiment design with non-normal response [J]. Industrial Engineering Journal，2008，11(2)：72－75. (in Chinese)

[6] Lewis S L，Montgomery D C，Myers R H. Confidence interval coverage for designed experiments analyzed with GLMs [J]. Journal of Quality Technology，2001，33(3)：279－292.

[7] Lewis S L，Montgomery D C，Myers R H. Examples of designed experiments with nonnormal responses [J]. Journal of Quality Technology，2001，33(3)：265－278.

[8] Lewis S L，Montgomery D C，Myers R H. The analysis of designed experiments with non-normal responses [J]. Quality Engineering，1999，12(2)：225－243.

[9] Weaver B P，Hamada M S. A Bayesian approach to the analysis of industrial experiments：An illustration with binomial count data [J]. Quality Engineering，2008，20(3)：269－280.

[10] Robinson T J，Anderson-Cook C M，Hamada M S. Bayesian analysis of split-plot experiments with nonnormal responses for evaluating nonstandard performance criteria [J]. Technometrics，2009，51(1)：56－65.

[11] 汪建均，马义中，汪 新. 两阶段的贝叶斯模型选择与筛选试验分析 [J]. 系统工程理论与实践，2011，31(8)：1447－1453.
Wang Jianjun，Ma Yizhong，Wang Xin. Two-stage Bayesian model choice and analysis of screening experiments [J]. System Engineering-Theory & Practice，2011，31(8)：1447－1453. (in Chinese)

[12] George E I，McCulloch R E. Variable selection via Gibbs sampling [J]. Journal of The American Statistical Association，1993，88(423)：881－889.

[13] 代逸生. 变结构线性回归模型显著性检验的贝叶斯方法 [J]. 管理科学学报，1999，2(1)：50－53.
Dai Yisheng. Bayesian method for testing of variable structure linear regression model [J]. Journal of Management Sciences in China，1999，2(1)：50－53. (in Chinese)

[14] 崔庆安，何 桢，崔 楠. 基于 SVM 的 RSM 模型拟合方法研究 [J]. 管理科学学报，2008，11(1)：31－41.
Cui Qingan，He Zhen，Cui Nan. SVM-based RSM model fitting approach [J]. Journal of Management Sciences in China，2008，11(1)：31－41. (in Chinese)

[15] Chipman H，Hamada M，Wu C F J. A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing [J]. Technometrics，1997，11(4)：372－381.

[16] Dellaportas P，Forster J J，Ntzoufras I. On Bayesian model and variable selection using MCMC [J]. Statistics and Computing，2002，12(1)：27－36.

[17] Ntzoufras I. Gibbs variable selection using BUGS [J]. Journal of Statistical Software，2002，7(7)：1－19.

[18] Ntzoufras I. Bayesian Modeling Using WinBUGS [M]. New Jersey：John Wiley & Sons Inc，2009.

[19] Ibrahim J G, Laud P W. On Bayesian analysis of generalized linear models using Jeffreys's prior [J]. Journal of The American Statistical Association, 1991, 86(416): 981 – 986.

[20] SAS INSTITUTE I. SAS/STAT ® 9.2 User's Guide [M]. Gary, NC: SAS Institute Inc, 2008.

[21] Barbieri M M, Berger J O. Optimal predictive model selection [J]. Annals of Statistics, 2004, 32(3): 870 – 897.

[22] Fouskakis D, Ntzoufras I, Draper D. Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care [J]. The Annals of Applied Statistics, 2009, 3(2): 663 – 690.

Bayesian variable and model selection based on generalized linear models

WANG Jian-jun, MA Yi-zhong

School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract: As for fractional factorial experiments with non-normal responses, a Bayesian variable and model selection approach based on generalized linear models (GLM) was proposed in the paper when the number of factors in screening experiments is large. Firstly, an empirical Bayesian prior was selected to consider the uncertainty of parameters in GLM. Secondly, we set a binary variable indicator for each variable in the linear predicator of GLM, and established a transformational relation between the variable indicators and the model indicators. Thirdly, we could identify significant factors and select the best model by the posterior probabilities of the variable indicators and the model indicators. Finally, a practical industrial example reveals that the proposed method can effectively identify significant factors in the fractional factorial experiment with non-normal responses.

Key words: Bayesian variable selection; fractional factorial experiment; generalized linear models; screening experiments; non-normal response