

# Pareto分布下屏蔽数据的贝叶斯统计分析及其应用 \*

李亚兰<sup>1</sup> 徐安察<sup>1,2\*</sup>

(<sup>1</sup>温州大学数学与信息科学学院, 温州, 325035; <sup>2</sup>西北工业大学理学院, 西安, 710072)

## 摘要

本文在Pareto分布下考虑了屏蔽数据的贝叶斯统计分析. 通过引入辅助变量来刻画失效的原因, 从而简化似然函数, 并使用贝叶斯、多层次贝叶斯以及经验贝叶斯三种方法来估计参数, 然后通过一个实例来比较三种方法的优劣, 最后讨论了避免先验分布中超参数选取的方法.

**关键词:** Pareto分布, 屏蔽数据, 贝叶斯分析.

**学科分类号:** O212.8.

## §1. 引言

考虑由 $J$ 个部件组成的串联系统, 用随机变量 $X_j$ 表示系统中第 $j$ 个部件的寿命,  $j = 1, 2, \dots, J$ . 假设 $X_j$ ,  $j = 1, 2, \dots, J$ 相互独立,  $X_j$ 的密度函数为 $f_j(\cdot|\boldsymbol{\theta}_j)$ , 其中 $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jd_j}) \in \boldsymbol{\Theta}_j \subseteq \mathbf{R}^{d_j}$ . 记 $X_j$ 的生存函数为 $R_j(t|\boldsymbol{\theta}_j) = \mathbb{P}(X_j > t|\boldsymbol{\theta}_j) = \int_t^\infty f_j(x|\boldsymbol{\theta}_j)dx$ . 在可靠性统计中, 生存函数也称可靠度函数. 此外, 假设参数 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_J$ 各不相同. 记 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_J)$ , 则 $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbf{R}^d$ , 其中 $d = \sum_{j=1}^J d_j$ . 记 $T = \min\{X_1, X_2, \dots, X_J\}$ , 则 $T$ 就为该串联系统的寿命. 我们用 $f_T(t|\boldsymbol{\theta})$ 和 $R_T(t|\boldsymbol{\theta})$ 分别表示 $T$ 的密度函数和可靠度函数. 当对串联系统做寿命试验时, 除了能够观测到系统的失效时间外, 在理想的状态下还能检测出是哪个部件导致系统失效. 记随机变量 $K$ 表示导致系统失效的部件的编号, 在 $\{1, 2, \dots, J\}$ 中取值.  $K = j$ 表示第 $j$ 个部件导致系统失效.  $X_j$ ,  $j = 1, 2, \dots, J$ 之间相互独立及密度的连续性保证了 $K$ 只能在 $\{1, 2, \dots, J\}$ 中取到单个值, 因为多个部件同时导致系统失效的概率几乎处处为0. 但是, 由于诸多原因(如缺少检测失效的设备, 时间和费用的限制, 检测给产品带来的破坏等), 精确的失效机理 $K$ 往往观测不到, 相反地, 只是观测到一些可能的失效机理(集合). 称这种集合为最小随机子集(minimum random set), 记为 $M$ , 详见Guess等(1991). 下面举一个例子来更直观地了解屏蔽数据的形式.

**例 1** 表1数据来自于Reiser等(1995), 该数据是对682台IBM电脑做逐次截尾试验得到的. 一台电脑可看成一个串联系统, 导致电脑失效的主要部件是三个部件: 主板(mother board)、光驱(disc drive)和电源(power supply), 分别记为部件1、2和3.

\*国家自然科学基金(11201345, 71401134)、中国博士后科学基金(2015M572598)和浙江省自然科学基金(LY15G010006)资助.

\*通讯作者, E-mail: xuancha@wzu.edu.cn.

本文2014年12月30日收到, 2015年1月13日收到修改稿.

doi: 10.3969/j.issn.1001-4268.2015.03.005

表1 IBM电脑部件失效数据\*

失效时间(小时)	1	1	1	1	16	17	21	222
失效机理	{1}	{1}	{1,3}	{1,2,3}	{3}	{2,3}	{2}	{2}

\* 348个未失效的系统在67小时时被终止试验, 246个未失效的系统在200小时时被终止试验, 26个未失效的系统在800小时时被终止试验及54个未失效的系统在4000小时时被终止试验.

表1所列的数据就是屏蔽数据. 例如第3个观测值, 其失效时间为1小时, 失效机理为{1,3}, 这说明不能确定是部件1还是部件3导致系统失效, 但是可以确定一定是部件1或者部件3导致系统失效.

记 $\mathcal{C}$ 为观测值是否为截尾数据:  $\mathcal{C} = 0$ 为截尾数据;  $\mathcal{C} = 1$ 则为失效数据. 假设对 $n$ 个串联系统进行试验, 观测到的数据为 $(t_1, M_1, \mathcal{C}_1), \dots, (t_n, M_n, \mathcal{C}_n)$ , 其中 $M_i \in \{1, 2, \dots, J\}$ ,  $i = 1, 2, \dots, n$ . 则可写出似然函数为

$$\begin{aligned} L(\boldsymbol{\theta}|(t_i, M_i, \mathcal{C}_i), i = 1, 2, \dots, n) \\ = \prod_{i=1}^n \left\{ \left[ \sum_{j \in M_i} \mathsf{P}(M = M_i | T = t_i, K = j) h_j(t_i | \boldsymbol{\theta}_j) \right]^{\mathcal{C}_i} \prod_{k=1}^J R_k(t_i | \boldsymbol{\theta}_k) \right\}. \end{aligned} \quad (1.1)$$

通常地, 称 $\mathsf{P}(M = M_0 | T = t, K = j)$ 为屏蔽概率. 若 $\forall j_1, j_2 \in M_0$ , 有

$$\mathsf{P}(M = M_0 | T = t, K = j_1) = \mathsf{P}(M = M_0 | T = t, K = j_2) \equiv p(M_0), \quad (1.2)$$

即, 屏蔽概率与系统的失效时间和失效机理都无关, 则似然函数(1.1)可简化为

$$\begin{aligned} L(\boldsymbol{\theta}|(t_i, M_i, \mathcal{C}_i), i = 1, 2, \dots, n) &= \prod_{i=1}^n \left\{ \left[ p(M_i) \sum_{j \in M_i} h_j(t_i | \boldsymbol{\theta}_j) \right]^{\mathcal{C}_i} \prod_{k=1}^J R_k(t_i | \boldsymbol{\theta}_k) \right\} \\ &\propto \prod_{i=1}^n \left\{ \left[ \sum_{j \in M_i} h_j(t_i | \boldsymbol{\theta}_j) \right]^{\mathcal{C}_i} \prod_{k=1}^J R_k(t_i | \boldsymbol{\theta}_k) \right\}. \end{aligned} \quad (1.3)$$

称屏蔽概率满足(1.2)的假设为对称性假设. 但是由于对称性假设太强, 后来有一些文献放宽了该假设. 例如, 下面的假设

$$\mathsf{P}(M = M_0 | T = t, K = j) = \mathsf{P}(M = M_0 | K = j) = p_j(M_0), \quad (1.4)$$

即, 屏蔽概率与系统的失效时间无关, 但与失效机理有关. 则这在一定程度上放宽了对称性假设. 需要指出的是, 屏蔽概率在(1.4)的假设下是有约束条件的. 记 $\mathcal{M}$ 为集合{1, 2, ..., J}所有非空子集的集合, 以及 $\mathcal{M}_j = \{M_0 \in \mathcal{M} : j \in M_0\}$ . 在 $K = j$ 的条件下, 由于 $M_0$ 只可能在 $\mathcal{M}_j$ 中取值, 因此 $p_j(M_0) = \mathsf{P}(M = M_0 | K = j) = 0, \forall M_0 \in \mathcal{M}_j^c = \mathcal{M} - \mathcal{M}_j$ . 在(1.4)的假设下, 屏蔽概率的约束条件为

$$\sum_{M_0 \in \mathcal{M}} p_j(M_0) = \sum_{M_0 \in \mathcal{M}_j} p_j(M_0) = 1, \quad j = 1, 2, \dots, J. \quad (1.5)$$

屏蔽概率的参数个数共有  $J(2^{J-1} - 1)$  个。虽然在  $J = 2$  时, 屏蔽概率的参数个数只有 2 个, 但  $J = 4$  时, 屏蔽概率的参数个数就有 28 个。因此当  $J > 2$  时, 估计模型中参数的难度将会大大增加。

对于屏蔽数据的问题, 最早由 Friedman 和 Gertsbakh (1980) 提出使用极大似然法进行分析。在上世纪八九十年代, 一些学者的主要工作是针对具体的分布讨论参数极大似然法的显式解及区间估计, 见 Mayakawa (1984), Usher 和 Hodgson (1988), Lin 等 (1993, 1996)。在贝叶斯方法方面, 也产生了很多成果。值得一提的是, Berger 和 Sun (1993) 针对 Poly-Weibull 分布提出了引入辅助变量的思想, 在等概率假设下, 很好地解决了完全屏蔽数据的统计分析问题。此后很多学者在 Berger 和 Sun (1993) 的基础上做了一些推广, 见 Mukhopadhyay 和 Basu (1997), Basu 等 (1999)。新世纪后, 很多学者的工作集中在怎样去掉等概率假设。Kuo 和 Yang (2000) 在假设有两个部件的串联系统并且部件的寿命服从指数或者威布尔分布时, 提出两种不同的屏蔽概率模型, 一种是假设屏蔽概率与时间无关, 另一种是假设屏蔽概率是时间的指数递减函数。在假设屏蔽概率与时间无关时, Basu 等 (2003) 提出了 一般的贝叶斯分析框架可以用来分析不同截尾类型的屏蔽数据; Mukhopadhyay (2006) 用辅助变量的思想提出 EM 算法得到参数的估计; Xu 和 Tang (2009) 讨论了屏蔽数据的信息损失以及超参数的选取对参数估计的影响。当屏蔽数据带有协变量信息时, Sen 等 (2010) 选择比例失效模型来刻画因变量与协变量之间关系, 并且用贝叶斯方法估计模型中的参数。Xu 等 (2014) 在步进应力加速寿命试验下提出多层次贝叶斯方法来分析屏蔽数据。

Xu 和 Tang (2009) 在  $J = 2$  时考虑了 Pareto 分布屏蔽数据的贝叶斯统计分析, 但是只用了贝叶斯方法来分析了屏蔽数据。本文将在假设 (1.4) 下对一般情况 ( $J > 2$ ) 讨论 Pareto 分布屏蔽数据的贝叶斯统计分析, 提出三种贝叶斯方法处理屏蔽数据, 并在实际数据分析中比较了这三种方法的优劣。第二节首先给出了 Pareto 分布下屏蔽数据的似然函数以及引入的辅助变量, 然后用贝叶斯、多层次贝叶斯以及经验贝叶斯等方法来估计模型中的参数, 并给出估计参数的 Gibbs 抽样步骤; 在第三节中用这三种方法分析了 Reiser 等 (1995) 的例子, 并比较了这三种方法的优劣; 最后讨论了避免先验分布中超参数选择的方法。

## §2. 贝叶斯统计分析

由于其厚尾性, Pareto 分布在金融保险行业应用比较广泛, 当然在可靠性统计当中也有一定的应用。Lindley 和 Singpurwalla (1986) 指出 Pareto 分布可以很好地刻画随机变化的环境。Pareto 分布的密度函数和可靠度函数分别具有如下形式:

$$f(t|\lambda, \tau) = \frac{\lambda}{\tau} \left(\frac{\tau}{t}\right)^{\lambda+1}, \quad t > \tau, \tau > 0; \quad R(t|\lambda, \tau) = \left(\frac{\tau}{t}\right)^\lambda. \quad (2.1)$$

我们将一个随机变量  $T$  服从 Pareto 分布简记为:  $T \sim \text{Pa}(\tau, \lambda)$ , 其中  $\tau$  为刻度参数,  $\lambda$  为形状参数。因此, 对一个  $J$  个部件的串联系统, 若第  $j$  个部件的寿命服从  $\text{Pa}(\tau_j, \lambda_j)$ , 且部件之间相

互独立, 则该系统在时刻  $t$  的可靠度函数和失效率函数就分别为

$$R(t) = \prod_{j=1}^J \left( \frac{\tau_j}{t} \right)^{\lambda_j}, \quad t > \max(\tau_1, \tau_2, \dots, \tau_J); \quad h(t) = \sum_{j=1}^J \frac{\lambda_j}{t}. \quad (2.2)$$

假设对  $n$  个串联系统进行试验, 观测数据为  $(t_i, M_i, \mathcal{C}_i), i = 1, 2, \dots, n$ , 由(1.1)可知, 似然函数为

$$\begin{aligned} L((\boldsymbol{\theta}, \mathbf{p}) | (t_i, M_i, \mathcal{C}_i), i = 1, 2, \dots, n) &= \prod_{i=1}^n \left\{ \left[ \sum_{j \in M_i} \lambda_j p_j(M_i) / t_i \right]^{\mathcal{C}_i} \prod_{j=1}^J \left( \frac{\tau_j}{t_i} \right)^{\lambda_j} \right\} \\ &\propto \exp \left\{ \sum_{j=1}^J \lambda_j (n \log \tau_j - lt) \right\} \prod_{i=1}^n \left[ \sum_{j \in M_i} \lambda_j p_j(M_i) \right]^{\mathcal{C}_i}, \end{aligned}$$

其中  $\mathbf{p}$  为相互不同的  $p_j(M_i), j = 1, 2, \dots, J, i = 1, 2, \dots, n$  的集合,  $lt = \sum_{i=1}^n \log t_i$ .

引入辅助变量  $Z_i = (Z_{i1}, Z_{i2}, \dots, Z_{iJ}), i = 1, 2, \dots, n$ , 其中  $Z_{ij} = 0, j \notin M_i; Z_{ij} = I(X_j = t_i), j \in M_i, I(\cdot)$  为示性函数. 则基于“完全数据”  $(\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}) = \{(t_i, Z_i, M_i, \mathcal{C}_i), i = 1, 2, \dots, n\}$ , 似然函数可简化为

$$\begin{aligned} L((\boldsymbol{\theta}, \mathbf{p}) | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C})) &\propto \exp \left\{ \sum_{j=1}^J \lambda_j (n \log \tau_j - lt) \right\} \prod_{i=1}^n \prod_{j \in M_i} [\lambda_j p_j(M_i)]^{Z_{ij} \mathcal{C}_i} \\ &= \exp \left\{ \sum_{j=1}^J \lambda_j (n \log \tau_j - lt) \right\} \prod_{i=1}^n \prod_{j=1}^J [\lambda_j p_j(M_i)]^{Z_{ij} \mathcal{C}_i}, \quad (2.3) \end{aligned}$$

其中定义  $0^0 = 1$ .

## 2.1 贝叶斯方法

由于 Gamma 分布是条件共轭先验, 因此对  $\lambda_j$ , 选取 Gamma 分布  $\Gamma(a_j, b_j)$  作为其先验. 对  $\tau_j$ , 我们选取两种先验: 无信息先验和条件共轭先验, 即,

$$\tau_j \propto 1/\tau_j \text{ 或 } \tau_j | \lambda_j \propto \lambda_j d_j^{-c_j \lambda_j} \tau_j^{c_j \lambda_j - 1}, \quad j = 1, 2, \dots, J.$$

由  $\sum_{M_0 \in \mathcal{M}_j} p_j(M_0) = 1$  且  $\mathcal{M}_j$  中共有  $2^{J-1}$  个包含  $j$  的集合, 我们给  $\mathbf{p}_j = \{p_j(M_0) : M_0 \in \mathcal{M}_j\}$  选取  $2^{J-1}$  维的 Dirichlet 先验, 即  $\mathbf{p}_j \sim D(\boldsymbol{\alpha}_j)$ , 其中参数  $\boldsymbol{\alpha}_j$  为  $2^{J-1}$  维的向量. 由(2.3)知, Dirichlet 先验为  $\mathbf{p}_j$  的共轭先验.

当  $\tau_j \propto 1/\tau_j, j = 1, 2, \dots, J$  时, 参数的满条件后验分布为

$$P(Z_i = e_j | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}) / Z_i, \psi) = \frac{\lambda_j p_j(M_i)}{\sum_{k \in M_i} \lambda_k p_k(M_i)}, \quad j \in M_i,$$

$$\mathbf{p}_j | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \psi / \mathbf{p}_j \sim D(\cdot), \quad (2.4)$$

$$\lambda_j | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \psi / \lambda_j \sim \Gamma \left( \sum_{i=1}^n Z_{ij} \mathcal{C}_i + a_j, lt - n \log \tau_j + b_j \right),$$

$$\tau_j | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \psi / \tau_j \sim GU(n \lambda_j - 1, 0, t_{(j1)}), \quad j = 1, 2, \dots, J,$$

其中  $\psi = (\theta, p)$ ,  $e_j$  为第  $j$  个元素为 1, 其他元素为 0 的  $J$  维向量;  $\text{GU}(a, b, c)$  表示参数为  $a, b$  和  $c$  的广义均匀分布, 其密度函数为  $(a+1)x^a/[c^{a+1} - b^{a+1}]$ ,  $b < x < c$ ,  $a > 0$ ,  $t_{(j1)}$  为观测到系统失效来自部件  $j$  时失效时间的最小次序统计量. 这里我们只知道  $p_j$  的后验满条件分布为 Dirichlet 分布, 具体参数的值要根据观测数据才能明确给出.

当  $\tau_j | \lambda_j \propto \lambda_j d_j^{-c_j \lambda_j} \tau_j^{c_j \lambda_j - 1}$ ,  $j = 1, 2, \dots, J$  时, 参数的满条件后验分布为

$$\begin{aligned} P(Z_i = e_j | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}) / Z_i, \psi) &= \frac{\lambda_j p_j(M_i)}{\sum_{k \in M_i} \lambda_k p_k(M_i)}, \quad j \in M_i, \\ p_j | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \psi / p_j &\sim D(\cdot), \\ \lambda_j | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \psi / \lambda_j &\sim \Gamma\left(\sum_{i=1}^n Z_{ij} \mathcal{C}_i + a_j + 1, \beta_j\right), \\ \tau_j | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \psi / \tau_j &\sim \text{GU}((n + c_j) \lambda_j - 1, 0, t_{(j1)}), \quad j = 1, 2, \dots, J, \end{aligned} \tag{2.5}$$

其中  $\beta_j = lt + c_j \log d_j + b_j - (n + c_j) \log \tau_j$ .

Gibbs 抽样步骤如下:

1. 给定  $\psi$  的初值;
2. 根据满条件后验分布(2.4)或(2.5)抽取参数的样本;
3. 重复第二步直到收敛.

## 2.2 多层贝叶斯方法

多层贝叶斯方法就是把先验中的超参数看做随机变量, 选取分布函数作为超参数的先验. 例如在贝叶斯方法中, 我们选取了 Gamma 分布  $\Gamma(a_j, b_j)$  作为  $\lambda_j$  的先验, 在多层次贝叶斯方法中, 还可以给  $a_j, b_j$  指定先验分布. 具体使用方法可见下一节的实例分析.

## 2.3 经验贝叶斯方法

经验贝叶斯方法其实是一种两步法. 在第一步中, 基于观测数据估计出先验中的超参数. 例如在贝叶斯方法中选取了  $\Gamma(a_j, b_j)$  作为  $\lambda_j$  的先验, 则可基于观测数据  $(t_i, M_i, \mathcal{C}_i)$ ,  $i = 1, 2, \dots, n$  用极大似然法、EM 算法等估计出  $(a_j, b_j)$ , 记为  $(\hat{a}_j, \hat{b}_j)$ . 这时  $\lambda_j$  的先验就被确定为  $\Gamma(\hat{a}_j, \hat{b}_j)$ . 在第二步代入确定好的先验分布进行贝叶斯统计分析.

## §3. 实例分析

这一节将对引言中的例子进行分析. 这个例子最早由 Reiser 等(1995) 提出. 他们对屏蔽概率作了对称性假设, 并且假定部件的寿命分布为指数分布, 选取无信息先验, 得到了贝叶斯后验均值的显式表达式. Basu 等(1999) 则假定部件的寿命分布为威布尔分布, 对刻度参数选取无信息先验, 形状参数选取离散先验, 最后用 Gibbs 抽样得到了参数的贝叶

斯估计. Basu等(2003)用贝叶斯因子比较了部件寿命取三种不同分布(指数分布、威布尔分布和对数正态分布)时对数据拟合的好坏, 最后得出部件的寿命分布取威布尔分布时最好. 假定 $P(M = M_0|T = t, K = j) = p_j(M_0)$ 且部件寿命分布为威布尔分布时, Mukhopadhyay (2006)通过EM算法对这批数据进行了分析.

Basu等(1999)和Basu等(2003)指出各部件的寿命分布具有厚尾性, 因此这节我们将使用Pareto分布对数据进行分析. 这里有三个部件, 所以屏蔽概率 $\mathbf{p}_1, \mathbf{p}_2$ 和 $\mathbf{p}_3$ 都是4维的向量. 对屏蔽概率的先验做如下假定:

$$\begin{aligned}\mathbf{p}_1 &= (p_1(\{1\}), p_1(\{1, 2\}), p_1(\{1, 3\}), p_1(\{1, 2, 3\})) \sim D(\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14}), \\ \mathbf{p}_2 &= (p_2(\{2\}), p_2(\{1, 2\}), p_2(\{2, 3\}), p_2(\{1, 2, 3\})) \sim D(\alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{24}), \\ \mathbf{p}_3 &= (p_3(\{3\}), p_3(\{1, 3\}), p_3(\{2, 3\}), p_3(\{1, 2, 3\})) \sim D(\alpha_{31}, \alpha_{32}, \alpha_{33}, \alpha_{34}).\end{aligned}$$

由于没有先验的信息, 对Dirichlet分布的参数都取相同的值, 但是我们取三个不同的值(0.05, 1和5)来考察超参数是否会影响Pareto分布中的参数估计. 观测数据中只有3个数据的失效机理屏蔽, 分别是第3, 4和6个观测, 所以只需3个辅助变量, 分别记为 $Z_1, Z_2$ 和 $Z_3$ . 由此可以得到 $\mathbf{p}_j$ 后验满条件分布为

$$\begin{aligned}\mathbf{p}_1|\mathbf{Z} &\sim D(\alpha_{11} + 2, \alpha_{12}, \alpha_{13} + Z_{11}, \alpha_{14} + Z_{21}), \\ \mathbf{p}_2|\mathbf{Z} &\sim D(\alpha_{21} + 2, \alpha_{22}, \alpha_{23} + Z_{32}, \alpha_{24} + Z_{22}), \\ \mathbf{p}_3|\mathbf{Z} &\sim D(\alpha_{31} + 1, \alpha_{32} + Z_{13}, \alpha_{33} + Z_{33}, \alpha_{34} + Z_{23}).\end{aligned}$$

在(2.4)和(2.5)中, 只需要将 $n = 682, t_{(11)} = 1, t_{(21)} = 21, t_{(31)} = 16$ 代入可以得到 $\tau_j$ 的满条件后验分布; 将 $\sum_{i=1}^n Z_{i1} \mathcal{C}_i = 2 + Z_{11} + Z_{21}, \sum_{i=1}^n Z_{i2} \mathcal{C}_i = 2 + Z_{22} + Z_{32}, \sum_{i=1}^n Z_{i3} \mathcal{C}_i = 1 + Z_{13} + Z_{23} + Z_{33}$ 代入可以得到 $\lambda_j$ 的满条件后验分布. 首先取定Gamma先验中超参数的值都为20. 表2列出了参数的后验均值. 这里我们只列出 $\tau_j$ 取无信息先验的结果, 因为两种先验所得到的结果没什么差异.

从表中可以发现, 发生变化的只是屏蔽概率的值, 这个很显然, 因为本身数据是重截尾的, 而且观测数据少, 先验信息的不同会影响后验估计, 但是有趣的是, Pareto分布中参数的后验均值基本上没变, 这是比较理想的, 因为估计系统可靠度函数时只与Pareto分布的参数有关. 在Dirichlet先验中超参数值取0.05的Gibbs抽样过程中, 我们记录了辅助变量的值, 并计算了观测数据屏蔽时该失效来自哪个部件的概率, 结果列于表3.

我们还计算出了三者的DIC值, 分别为 $DIC_{0.05} = 355.62$ ,  $DIC_1 = 354.81$ 和 $DIC_5 = 354.12$ , 基本没什么差异. 此外, 在Dirichlet先验中超参数值取0.05时, Gamma先验的超参数取 $a_j = 15, b_j = 1$ , 计算了参数的后验均值, 结果列在表2中的 $(a_j = 15, b_j = 1)$ 行中, 可以发现屏蔽概率的估计没什么变化, 这说明Gamma先验的超参数选取对屏蔽概率的估计可能不会有影响; 但是其他参数的估计有一些改变, 即Gamma先验的超参数会影

表2 IBM数据各参数的后验均值

超参数( $\alpha$ )或方法	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\tau_1$	$\tau_2$	$\tau_3$
0.05	0.00614	0.0162	0.0139	0.7442	19.002	14.249
1	0.00614	0.0162	0.0140	0.7435	18.938	14.246
5	0.00615	0.0163	0.0139	0.7491	18.979	14.221
$(a_j = 15, b_j = 1)$	0.00561	0.0148	0.0128	0.726	18.808	14.044
经验贝叶斯	0.00468	0.00616	0.00897	0.6664	15.659	13.276
多层贝叶斯	0.02133	0.02792	0.02708	0.9302	19.898	15.151
超参数( $\alpha$ )或方法	$p_1(\{1\})$	$p_1(\{1, 2\})$	$p_1(\{1, 3\})$	$p_1(\{1, 2, 3\})$	$p_2(\{2\})$	$p_2(\{1, 2\})$
0.05	0.8048	0.0217	0.1072	0.0663	0.6762	0.0192
1	0.4698	0.1554	0.1947	0.1801	0.4345	0.1447
5	0.3109	0.2238	0.2355	0.2298	0.3048	0.2193
$(a_j = 15, b_j = 1)$	0.7986	0.0205	0.1074	0.0735	0.7133	0.0176
经验贝叶斯	0.7847	0.0183	0.1141	0.0829	0.7512	0.0164
多层贝叶斯	0.7507	0.0190	0.1302	0.1001	0.7118	0.0164
超参数( $\alpha$ )或方法	$p_2(\{2, 3\})$	$p_2(\{1, 2, 3\})$	$p_3(\{3\})$	$p_3(\{1, 3\})$	$p_3(\{2, 3\})$	$p_3(\{1, 2, 3\})$
0.05	0.1740	0.1306	0.4168	0.2709	0.1548	0.1575
1	0.2206	0.2002	0.3108	0.2632	0.2196	0.2064
5	0.2402	0.2357	0.2674	0.2546	0.2417	0.2363
$(a_j = 15, b_j = 1)$	0.1589	0.1102	0.3892	0.2595	0.1852	0.1661
经验贝叶斯	0.1315	0.1009	0.3742	0.2399	0.2142	0.1717
多层贝叶斯	0.1452	0.1266	0.4152	0.2435	0.2110	0.1303

表3  $(p_1, p_2, p_3) : p_j$  为失效来自第j部件的概率

第3个观测	第4个观测	第6个观测
(0.3433, 0, 0.6567)	(0.2386, 0.389, 0.3725)	(0, 0.4580, 0.5420)

响Pareto参数的估计, 进而影响系统可靠度函数的估计. 为此我们用经验贝叶斯方法来估计Gamma先验中的超参数. 首先由表3假设第3, 4和6个观测分别来自部件3, 2和3的失效, 然后用经验贝叶斯的方法可以估计出  $a_1 = 19.82$ ,  $b_1 = 1061.52$ ,  $a_2 = 24.89$ ,  $b_2 = 2896.32$ ,  $a_3 = 28.21$ ,  $b_3 = 1872.51$ . 模型中参数的估计列于表2的“经验贝叶斯”行中, 且DIC的值为246.58. 从DIC准则看, 经验贝叶斯方法所得到的先验更好. 但有趣的是, 经验贝叶斯所得结果的实际解释却非常不合理. 由参数的估计, 可以计算出系统的第一四分位寿命为1617.134年! 从实际情况看, 这是不可能的. 一般来说, 经验贝叶斯可以看作是选取超参数

值的一种客观方法, 但是在这个例子中, 由于观测数据较少而且是重截尾型的, 所以会产生上述的结果. 那么, 对这种类型的数据, 最好就是有一些可靠的信息, 比如说部件的第一四分位寿命大概为10年(86400小时), 则就有

$$1 - \left( \frac{\tau_j}{86400} \right)^{\lambda_j} = 0.25 \Rightarrow \lambda_j = \log 0.75 / (\log \tau_j - \log 86400), \quad j = 1, 2, 3.$$

由于  $\tau_j \leq t_{(j1)}$ , 可知  $\lambda_j \leq \log 0.75 / (\log t_{(j1)} - \log 86400)$ ,  $j = 1, 2, 3$ . 由之前得到的  $\tau_j$  的估计来看,  $\tau_j \geq t_{(j1)}/2$ . 这里我们取其下界为  $\tau_j \geq t_{(j1)}/10$ . 则  $\tau_j, j = 1, 2, 3$  就有约束条件, 具体为

$$0.02104 \leq \lambda_1 \leq 0.02530, \quad 0.02707 \leq \lambda_2 \leq 0.03456, \quad 0.02640 \leq \lambda_3 \leq 0.03347.$$

在Gamma先验中, 我们取超参数  $a_1, a_2$  和  $a_3$  的值为经验贝叶斯所得到的值, 那么根据  $\lambda_j$  的信息, 超参数  $b_j$  可以得到一个取值范围, 即

$$783.12 \leq b_1 \leq 941.79, \quad 720.05 \leq b_2 \leq 919.26, \quad 842.74 \leq b_3 \leq 1068.56.$$

考虑用多层次贝叶斯的方法来分析这个问题, 选取均匀分布作为超参数  $b_j$  的先验, 即

$$b_1 \sim U(783.12, 941.79), \quad b_2 \sim U(720.05, 919.26), \quad b_3 \sim U(842.74, 1068.56).$$

这时 Gibbs 抽样中需加入对  $b_j$  的抽样, 则可推得  $b_j$  和  $\lambda_j$  的满条件后验分布为

$$b_1 | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \boldsymbol{\lambda} \sim \Gamma(a_1 + 1, \lambda_1) \cdot I(783.12 \leq b_1 \leq 941.79),$$

$$b_2 | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \boldsymbol{\lambda} \sim \Gamma(a_2 + 1, \lambda_2) \cdot I(720.05 \leq b_2 \leq 919.26),$$

$$b_3 | (\mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \boldsymbol{\lambda} \sim \Gamma(a_3 + 1, \lambda_3) \cdot I(842.74 \leq b_3 \leq 1068.56),$$

$$\lambda_1 | (b_1, \mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \boldsymbol{\lambda} / \lambda_1 \sim \Gamma\left(\sum_{i=1}^n Z_{i1} \mathcal{C}_i + a_1, lt - n \log \tau_1 + b_1\right) \cdot I(0.02104 \leq \lambda_1 \leq 0.02530),$$

$$\lambda_2 | (b_2, \mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \boldsymbol{\lambda} / \lambda_2 \sim \Gamma\left(\sum_{i=1}^n Z_{i2} \mathcal{C}_i + a_2, lt - n \log \tau_2 + b_2\right) \cdot I(0.02707 \leq \lambda_2 \leq 0.03456),$$

$$\lambda_3 | (b_3, \mathbf{t}, \mathbf{Z}, \mathbf{M}, \mathcal{C}), \boldsymbol{\lambda} / \lambda_3 \sim \Gamma\left(\sum_{i=1}^n Z_{i3} \mathcal{C}_i + a_3, lt - n \log \tau_3 + b_3\right) \cdot I(0.02640 \leq \lambda_3 \leq 0.03347).$$

屏蔽概率和  $\tau_j$  的满条件后验分布不变. 模型中参数的估计列于表2的“多层次贝叶斯”行中, 且 DIC 的值为 424.62. 这时计算出系统的中位寿命为 7.79 年(67315 小时). 这个结果也同 Basu 等(2003) 所得到的结果相近. 相比于经验贝叶斯的结果, 7.79 年是一个更为实际的结果. 在分析数据时, 经验贝叶斯方法会更侧重于数据拟合, 这一点跟极大似然法相似, 而对于重截尾的数据类型, 这种处理方法往往会导致不合理的结果. 通过这个例子, 我们想说明的是对重截尾的数据类型多层次贝叶斯方法会更加有效, 由于加入一些可靠的信息而会使得结果更加合理.

## §4. 讨 论

本文考虑了Pareto模型下屏蔽数据的贝叶斯分析, 使用不同的贝叶斯方法对部件数  $J > 2$  的情况做了讨论。在贝叶斯方法中使用了两种不同的先验, 并给出了估计参数的 Gibbs 抽样步骤, 然后通过一个实例的具体分析说明了对于重截尾的数据, 多层贝叶斯的方法要比经验贝叶斯方法更好, 得到的结果更合理。

使用贝叶斯方法会涉及到超参数的敏感性问题。在计算中一般需要使用不同的超参数值进行计算, 然后看结果是否会出现较大的差异。当然如果有实际经验可参考的话, 超参数的值比较容易确定, 但是这方面的经验往往很难得到。在实例的分析中, 我们用了经验贝叶斯方法和多层贝叶斯方法。这两种方法可以从一定程度上解决超参数的选取问题。一般情况下, 这两种方法得到的结果会比较接近, 但是对样本量较小的情况, 我们推荐使用多层贝叶斯的方法。还有一种方法可以避免超参数的选取问题: 通过客观贝叶斯方法选取一种无信息先验, 例如Jeffreys先验和reference先验。对  $\text{Pa}(\tau, \lambda)$ , Jeffreys先验为  $\sqrt{1/\lambda - 1}/\tau$ 。根据讨厌参数的不同, reference先验会不同。当  $\lambda$  和  $\tau$  分别为讨厌参数时, 对应的reference先验分别为

$$\frac{1}{\tau\lambda} \quad \text{和} \quad \frac{\sqrt{1-\lambda}}{\tau}.$$

可知三种无信息先验都不包含超参数, 从而避免了超参数的选取问题。

## 参 考 文 献

- [1] Guess, F.M., Usher, J.S. and Hodgson, T.J., Estimating system and component reliabilities under partial information on cause of failure, *Journal of Statistical Planning and Inference*, **29(1-2)**(1991), 75–85.
- [2] Reiser, B., Guttman, I., Lin, D.K.J., Guess, F.M. and Usher, J.S., Bayesian inference for masked system lifetime data, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **44(1)**(1995), 79–90.
- [3] Friedman, L. and Gertsbakh, I.B., Maximum likelihood estimation in a minimum-type model with exponential and Weibull failure modes, *Journal of the American Statistical Association*, **75(370)**(1980), 460–465.
- [4] Miyakawa, M., Analysis of incomplete data in competing risks model, *IEEE Transactions on Reliability*, **33(4)**(1984), 293–296.
- [5] Usher, J.S. and Hodgson, T.J., Maximum likelihood analysis of component reliability using masked system life-test data, *IEEE Transactions on Reliability*, **37(5)**(1988), 550–555.
- [6] Lin, D.K.J., Usher, J.S. and Guess, F.M., Exact maximum likelihood estimation using masked system data, *IEEE Transactions on Reliability*, **42(4)**(1993), 631–635.
- [7] Lin, D.K.J., Usher, J.S. and Guess, F.M., Bayes estimation of component-reliability from masked system-life data, *IEEE Transactions on Reliability*, **45(2)**(1996), 233–237.
- [8] Berger, J.O. and Sun, D., Bayesian analysis for the Poly-Weibull distribution, *Journal of the American Statistical Association*, **88(424)**(1993), 1412–1418.

- [9] Mukhopadhyay, C. and Basu, A.P., Bayesian analysis of incomplete time and cause of failure data, *Journal of Statistical Planning and Inference*, **59**(1)(1997), 79–100.
- [10] Basu, S., Basu, A.P. and Mukhopadhyay, C., Bayesian analysis for masked system failure data using non-identical Weibull models, *Journal of Statistical Planning and Inference*, **78**(1-2)(1999), 255–275.
- [11] Kuo, L. and Yang, T.Y., Bayesian reliability modeling for masked system lifetime data, *Statistics and Probability Letters*, **47**(3)(2000), 229–241.
- [12] Basu, S., Sen, A. and Banerjee, M., Bayesian analysis of competing risks with partially masked cause of failure, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**(1)(2003), 77–93.
- [13] Mukhopadhyay, C., Maximum likelihood analysis of masked series system lifetime data, *Journal of Statistical Planning and Inference*, **136**(3)(2006), 803–838.
- [14] Xu, A. and Tang, Y., Bayesian analysis of Pareto reliability with dependent masked data, *IEEE Transactions on Reliability*, **58**(4)(2009), 583–588.
- [15] Sen, A., Banerjee, M., Li, Y. and Noone, A.M., A Bayesian approach to competing risks analysis with masked cause of death, *Statistics in Medicine*, **29**(16)(2010), 1681–1695.
- [16] Xu, A., Basu, S. and Tang, Y., A full Bayesian approach for masked data in step-stress accelerated life testing, *IEEE Transactions on Reliability*, **63**(3)(2014), 798–806.
- [17] Lindley, D.V. and Singpurwalla, N.D., Multivariate distributions for the life lengths of components of a system sharing a common environment, *Journal of Applied Probability*, **23**(2)(1986), 418–431.

## Bayesian Statistical Analysis and Application of Masked Data based on Pareto Distribution

LI YALAN<sup>1</sup>      XU ANCHA<sup>1,2</sup>

<sup>(1)</sup>*College of Mathematics and Information Science, Wenzhou University, Wenzhou, 325035*

<sup>(2)</sup>*College of Science, Northwestern Polytechnical University, Xi'an, 710072*)

In this paper Bayesian statistical analysis of masked data is considered based on the Pareto distribution. The likelihood function is simplified by introducing auxiliary variables, which describe the causes of failure. Three Bayesian approaches (Bayes using subjective priors, hierarchical Bayes and empirical Bayes) are utilized to estimate the parameters, and we compare these methods by analyzing a real data. Finally we discuss the method of avoiding the choice of the hyperparameters in the prior distributions.

**Keywords:** Pareto distribution, masked data, Bayesian analysis.

**AMS Subject Classification:** 62N05.