左截断右删失下正态分布参数单变点的 bayes 估计

李清宇[1] 黄介武[1]
LI Qingyu HUANG Jiewu

摘 要 针对随机变量的寿命试验数据服从正态分布且在缺失数据模型下的情况，对其参数识别以及估计，并应用到实际生活。在此研究中，首先构建正态分布的似然函数，考虑在左截断右删失数据模型下，使用逆变换法和筛选法处理缺失数据。然后将缺失数据模型下的似然函数转化为完全数据情况下的似然函数，可获得变点的位置及其他参数的满条件分布。最后，采用 Gibbs 抽样和 Metropolis-Hastings 算法相结合的 MCMC 方法获得各参数的 Gibbs 样本，并将其均值作为各参数的贝叶斯估计。结果显示：在相对误差、MC 误差的准则下，各参数的贝叶斯估计精度都较高。

关键词 正态分布；逆变换法；筛选法；Gibbs 抽样；贝叶斯估计

doi：10.3969/j.issn.1672-9528.2024.02.028

0 引言

变点问题是一个长久并且受到广泛关注的主题，几乎所有的学者都认同最早关于连续性试验方案的论文是由 Page[1] 在 1954 年发表于 Biornetrika 杂志上的。关于变点问题的分析可以参看文献 [2-6]。贝叶斯方法因在处理不确定性和复杂数据结构方面的优势受到关注，特别是存在截断删失数据的情况下，贝叶斯方法有更强的适应性。正态分布是一种应用很广泛的连续型分布。例如，电灯丝设备（如白炽灯泡和面包机加热元件）、集成电路中电线的强度等是服从正态分布的示例。有关正态分布的研究资料可以参看文献 [7-9]。在寿命试验中，左侧截断和右侧删失等问题经常出现，导致收集的数据不完整，对于这种类型的数据，在变点模型下的分析也备受关注。对左截断右删失数据下变点模型的研究可参看文献 [10-12]。何朝兵等人 [13-18] 也对在此数据类型下的其他分布进行了研究。

1 左截断右删失模型

设 (X, Y, T) 是连续性随机变量，变量 X 的分布函数为 $F(x;\lambda)=P(X\le x)$，密度函数为 $f(x;\lambda)$，在这种情况下，λ 表示未知参数 (或参数向量)；变量 Y 是右删失随机变量，分布函数为 $G(y)$，密度函数为 $g(y)$；变量 T 是左截断随机变量，分布函数为 $H(t)$，密度函数为 $h(t)$。假设 X, Y, T 是相互独立且取值都为非负的随机变量。对 n 个试验产品，左截断右删失模型只有当 $Z_i\ge T_i$ 时，可得到观察数据 (Z_i, T_i, δ_i)，而在 $Z_i< T_i$ 下无法得到任何观察值，其中引入变量：

$$Z_i = X_i \wedge Y_i = \min(X_i, Y_i), \delta_i = I(X_i \le Y_i) \quad (i = 1, 2, \cdots, n)$$

下面求样本的似然函数。

$$P(Z_i<T_i)=1-P(X_i\ge T_i, Y_i\ge T_i)=1-\int_0^{\infty}h(x)\overline{F}(x)\overline{G}(x)dx=u(\lambda) \quad (1)$$

式中：$\overline{F}=1-F, \overline{G}=1-G$。为了研究方便，引入示性变量 $v_i=I(\min(X_i,Y_i)\ge T_i), i=1,2,\cdots,n$。

基于观察数据 $\{(z_i,t_i,\delta_i): v_i=1, 1\le i\le n\}$ 的似然函数为：

$$\begin{aligned}L(\lambda)&=\prod_{i=1}^{n}\left\{\left[f(z_i;\lambda)\overline{G}(z_i)h(t_i)\right]^{v_i\delta_i}\left[g(z_i)\overline{F}(z_i;\lambda)h(t_i)\right]^{v_i(1-\delta_i)}\left[u(\lambda)\right]^{1-v_i}\right\}\\&=A\prod_{i=1}^{n}\left\{\left[f(z_i;\lambda)\right]^{v_i\delta_i}\left[\overline{F}(z_i;\lambda)h(t_i)\right]^{v_i(1-\delta_i)}\left[u(\lambda)\right]^{1-v_i}\right\}\end{aligned} \quad (2)$$

式中：$A=\left[h(t_i)\right]\sum_{i=1}^{n}v_i\prod_{i=1}^{n}\left\{\left[\overline{G}(z_i)\right]^{v_i}\left[g(z_i)\right]^{v_i(1-\delta_i)}\right\}$。

2 正态分布的完全数据似然函数

若 x 的密度为 $f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty<x<\infty, \sigma>0$，则称 X 服从均值为 μ 方差为 σ^2 的正态分布，记 $X\sim N(\mu,\sigma^2)$。在左截断右删失数据模型下，设产品寿命 $X\sim N(\mu,\sigma^2)$，令 δ 表示 δ_i 组成的向量，z 表示 z_i 组成的向量，记 $v=(1,\cdots,v_n)$。

$$\begin{aligned}L(\lambda)&=\prod_{i=1}^{n}\left\{\left[f(z_i;\lambda)\overline{G}(z_i)h(t_i)\right]^{v_i\delta_i}\left[g(z_i)\overline{F}(z_i;\lambda)h(t_i)\right]^{v_i(1-\delta_i)}\left[u(\lambda)\right]^{1-v_i}\right\}\\&=A\prod_{i=1}^{n}\left\{\left[f(z_i;\lambda)\right]^{v_i\delta_i}\left[\overline{F}(z_i;\lambda)h(t_i)\right]^{v_i(1-\delta_i)}\left[u(\lambda)\right]^{1-v_i}\right\}\end{aligned} \quad (3)$$

由于左截断右删失数据 X_i 在 $v_i=0$ 或 $v_i=1$、$\delta_i=0$ 时缺失，则需分别考虑 $v_i=0$ 或者 $v_i=1$、$\delta_i=0$ 两种情况，通过添加数据

1. 贵州民族大学数据科学与信息工程学院 贵州贵阳 550025

[基金项目] 贵州省教育厅自然科学研究项目（编号：黔教技〔2022〕015 号）；贵州省高等学校大数据分析与智能计算重点实验室（编号：黔教技〔2023〕012 号）

就可以得到完全数据的似然函数，具体如下。

（1）当 v_i=1、δ_i=0 时，添加一个数据 Z_{1i}=X_{1i}=z_{1i}。在 $X_i > z_i$ 的条件下，z_{1i} 的条件密度为：

$$\psi_1(x;\lambda,z_i)=\frac{f(x;\lambda)}{1-F(z_i;\lambda)}(x>z_i) \tag{4}$$

Z_{1i} 条件分布是区间 $(z_i, +\infty)$ 上的截断正态分布 $N(\mu, \sigma^2)$，其样本可以表示为：

$$z_{1i}=F^{-1}\left\{F(z_i)+U\left[1-F(z_i)\right]\right\} \tag{5}$$

式中：F 是 $N(\mu, \sigma^2)$ 的分布函数，F^{-1} 为反函数，U 为均匀分布 $U(0,1)$ 的随机样本。在参数已知且符合逆变换法条件的情况下，利用逆变换法实现对 z_{1i} 的随机生成。

（2）当 v_i=0 时，添加一个数据 Z_{2i}=X_{1i}=z_{2i}。在 v_i=0 时，即 $\min(X_i < Y_i) < T_i$ 的条件下，Z_{2i} 的条件密度函数为：

$$\begin{aligned}\psi_2(x;\lambda)&=\left[u(\lambda)\right]^{-1}f(x;\lambda)P(\{Y>x,T>x\}\cup\{Y\le x,T>Y\})\\&=\left[u(\lambda)\right]^{-1}f(x;\lambda)\left[\overline{G}(x)\overline{H}(x)+\int_0^x g(x)\overline{H}(x)dx\right]\\&\le\left[u(\lambda)\right]^{-1}f(x;\lambda)\end{aligned} \tag{6}$$

由 $f(x;\lambda)$ 是 x_i 的密度函数，X_i 与 Z_{2i} 在有相同的取值且符合筛选抽样条件的情况下，利用筛选法随机生成 Z_{2i} 的值 z_{2i}，z_{2i} 抽样的具体步骤如下。

①从均匀分布 $U(0,1)$ 中抽取 u，从 $f(x;\lambda)$ 中抽取 x。

②若 $u \le \overline{G}(x)\overline{H}(x)+\int_0^x g(x)\overline{H}(x)dx$，则 z_{2i}=x，结束，否则的话回到步骤①。

令 u_1 代表 z_{1i} 构成的向量，u_2 代表 z_{2i} 构成的向量，则完全数据似然函数为：

$$\begin{aligned}L(z,u_1,u_2,v,\delta|\lambda)&\propto\prod_{i=1}^{n}\left\{\left[f(z_i;\lambda)\right]^{v_i\delta_i}\left[f(z_{1i};\lambda)\right]^{v_i(1-\delta)}\left[f(z_{2i};\lambda)\right]^{1-v_i}\right\}\\&=\left(\frac{1}{2}\pi\sigma\right)^n\left[\prod_{i=1}^{n}z_i^{v_i\delta_i}z_{1i}^{v_i(1-\delta)}z_{2i}^{1-v_i}\right]^{-1}e^{-\frac{s}{2\sigma^2}}\\&\propto\left(\frac{1}{\sigma}\right)^n e^{-\frac{s}{2\sigma^2}}\end{aligned} \tag{7}$$

式中：$s=\sum_{i=1}^{n}\left[v_i\delta_i(z_i-\mu)^2+v_i(1-\delta_i)(z_{1i}-\mu)^2+(1-v_i)(z_{1i}-\mu)^2\right]$

3 正态分布参数单变点的贝叶斯估计

正态分布参数单变点模型为：

$$X_i\sim\begin{cases}N(\mu_1,\sigma_1^2)\,i=1,\cdots,k\\N(\mu_2,\sigma_2^2)\,i=k+1,\cdots,n\end{cases} \tag{8}$$

式中：$(\mu_1, \sigma_1^2)(\mu_2, \sigma_2^2)$ 互不相等，$1 \le k \le n$-1，则此称为参数单变点模型。以下是对变点位置 k 和参数 μ_1、μ_2、σ_1^2、σ_2^2 进行贝叶斯估计的具体过程，记 β=$(k, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$，则此变点问题的似然函数为：

$$L(z,u_1,u_2,v,\delta|\beta)\propto\left[\left(\frac{1}{\sigma_1}\right)^k e^{-\frac{s_1}{2\sigma_1^2}}\right]\left[\left(\frac{1}{\sigma_2}\right)^{n-k}e^{-\frac{s_2}{2\sigma_2^2}}\right] \tag{9}$$

式中：

$$\begin{aligned}s_1&=\sum_{i=1}^{k}\left[v_i\delta_i(z_i-\mu_1)+v_i(1-\delta_i)(z_{1i}-\mu_1)^2+(1-v_i)(z_{2i}-\mu_1)^2\right]\\s_2&=\sum_{i=k+1}^{n}\left[v_i\delta_i(z_i-\mu_2)+v_i(1-\delta_i)(z_{1i}-\mu_2)^2+(1-v_i)(z_{2i}-\mu_2)^2\right]\end{aligned} \tag{10}$$

下面讨论各参数的先验分布。

（1）对于 k 取无信息先验分布，可取公式为：

$$\pi(k)=\frac{1}{C_n^1}=\frac{1}{n}(1\le k\le n) \tag{11}$$

式中：$\pi(k)$ 实际上是一种无信息先验分布。

（2）对于 (μ_i, σ_i^2) 的先验分布取值，如果无先验信息，按照贝叶斯假设取无信息先验分布 $\pi(\mu_i,\sigma_i^2)\propto 1,\mu_i>0,\sigma_i>0,i=1,2$，根据先验信息，如果 μ_i 与 σ_i^2 独立，μ_i 取共轭先验正态分布，σ_i^2 取共轭先验伽玛分布，这时它们不再是共轭的先验分布。根据先验信息，如果 μ_i 与 σ_i^2 不独立，(μ_i, σ_i^2) 的先验分布取共轭先验分布正态 - 伽玛分布。先验分布中的超参数可以利用先验矩或先验分位数等方法进行确定，充分利用先验信息，假设通过调节超参数可以使用更加合理的先验分布。

假设现在没有先验信息，所以 k 与 (μ_i, σ_i^2) 都可以取无信息先验分布，即：

$$\pi(k)=\frac{1}{C_n^1}=\frac{1}{n}(1\le k\le n),\pi(\mu_i,\sigma_i^2)\propto 1,\mu_i>0,\sigma_i>0,i=1,2. \tag{12}$$

假设 k、(μ_1, σ_1^2)、(μ_2, σ_2^2) 相互独立，则：

$$\begin{aligned}L(\beta|z,u_1,u_2,v,\delta)&\propto\pi(k)\pi(\mu_1,\sigma_1^2)\pi(\mu_2,\sigma_2^2)L(z,u_1,u_2,v,\delta|\beta)\\&\propto L(z,u_1,u_2,v,\delta|\beta)\end{aligned} \tag{13}$$

式中：$\lambda_i=(\mu_i, \sigma_i^2)$，$i=1, 2$。

当 v_i=1、δ_i=0 时，则：

$$\pi(z_{1i}|\beta,z,z_{-1i},u_2,v,\delta)\propto\varphi_1(z_{1i};\beta,z_i)=\begin{cases}\psi_1(z_{1i};\lambda_1,z_i)(i=1,2,\cdots,k)\\\psi_1(z_{1i};\lambda_2,z_i)(i=k+1,\cdots,n)\end{cases} \tag{14}$$

式中：$z_{-1i}=\{z_{1j}:j\ne i\}$。

当 v_i=0 时，则：

$$\pi(z_{2i}|\beta,z,z_{-2i},u_1,v,\delta)\propto\varphi_2(z_{2i};\beta,z_i)=\begin{cases}\psi_2(z_{2i};\lambda_1)(i=1,2,\cdots,k)\\\psi_2(z_{2i};\lambda_2)(i=k+1,\cdots,n)\end{cases} \tag{15}$$

式中：$z_{-2i}=\{z_{2j}:j\ne i\}$。满条件分布中的“条件”用“•”表示，例如 $\pi(\mu_1|k, \mu_2, \sigma_1^2, \sigma_2^2, z, u_1, u_2, v, \delta)$ 记为 $\pi(\mu_1|\bullet)$。

下面对各参数的满条件分布进行求解。

$$\begin{aligned}\pi(\mu_1|\bullet)&\propto e^{-\frac{s_1}{2\sigma_1^2}}\\&=\exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{k_1}\left[v_i\delta_i(z_i-\mu_1)^2+v_i(1-\delta_i)(z_{1i}-\mu_1)^2+(1-v_i)(z_{2i}-\mu_1)^2\right]\right\}\\&\propto\exp\left\{-\frac{1}{2\sigma_1^2}\left(\mu_1^2-2\overline{d_1}\mu_1\right)\right\}\end{aligned} \tag{16}$$

式中：$\overline{d_1}=\frac{1}{k}\sum_{i=1}^{k}\left[v_i\delta_i z_i+v_i(1-\delta_i)z_{1i}+(1-v_i)z_{2i}\right]$，记 $\overline{\sigma_1^2}=\frac{\sigma_1^2}{k}$。同理可得：

$$\pi(\mu_2|\bullet)\propto\exp\left\{-\frac{1}{2\overline{\sigma_2^2}}\left(\mu_2-\overline{d_2}\right)^2\right\}\propto N\left(\overline{d_2},\overline{\sigma_2^2}\right) \tag{17}$$

式中：

$$\overline{d_2}=\frac{1}{n-k_1}\sum_{i=k+1}^{n}\left[v_i\delta_i z_i+v_i\left(1-\delta_i\right)z_{1i}+\left(1-v_i\right)z_{2i}\right],\overline{\sigma_2^2}=\frac{\sigma_2^2}{(n-k)} \tag{18}$$

下面求 σ_1^2、σ_2^2 的满条件分布。

$$\begin{aligned}&\pi\left(\sigma_1^2|\bullet\right)\propto\left(\frac{1}{\sigma_1^2}\right)^k\exp\left(-\frac{s_1}{2\sigma_1^2}\right)\propto IGa\left(k-1,\frac{s_1}{2}\right)\\&\pi\left(\sigma_2^2|\bullet\right)\propto\left(\frac{1}{\sigma_2^2}\right)^{n-k}\exp\left(-\frac{s_2}{2\sigma_2^2}\right)\propto IGa\left(n-k-1,\frac{s_2}{2}\right)\end{aligned} \tag{19}$$

k 的满条件分布为：

$$\pi\left(k|\bullet\right)\propto\left(\frac{\sigma_2^2}{\sigma_1^2}\right)^k\exp\left(-\left(\frac{s_1}{2\sigma_1^2}+\frac{s_2}{2\sigma_2^2}\right)\right)\left(1\le k\le n\right) \tag{20}$$

由上文可知各参数的满条件分布，则利用 MCMC 方法在后验分布的稳定状态下得到参数的概率分布。$\mu_1,\mu_2,\sigma_1^2,\sigma_2^2$ 的满条件分布为标准形式，则采用 Gibbs 抽样；对于 k 选择离散型均匀分布。下面详细描述 MCMC 方法的步骤。

首先，确定一个初始点 $\beta^{(0)}=(k^{(0)},\mu_1^{(0)},\mu_2^{(0)},\sigma_1^{2(0)},\sigma_2^{2(0)})$，第 t 次迭代估计为 $\beta^{(t-1)}$，第 t 次迭代分为以下几步。

（1）当 v_i=1、δ_i=0 时，从分布 $\varphi_1\left(z_{1i};\beta^{(t-1)},z_i\right)$ 抽 $z_{1i}^{(t)}$，令 $u_1^{(t)}$ 表示 $z_{1i}^{(t)}$ 组成的向量。

（2）当 v_i=0 时，从分布 $\varphi_2\left(z_{2i};\beta^{(t-1)}\right)$ 抽 $z_{2i}^{(t)}$，令 $u_2^{(t)}$ 表示 $z_{2i}^{(t)}$ 组成的向量。

（3）从 $\pi(\mu_1|k^{(t-1)},\mu_2^{(t-1)},\sigma_1^{2(t-1)},\sigma_2^{2(t-1)},z,u_1^{(t)},u_2^{(t)},v,\delta)$ 抽 $\mu_1^{(t)}$。

（4）从 $\pi(\mu_2|k^{(t-1)},\mu_1^{(t-1)},\sigma_1^{2(t-1)},\sigma_2^{2(t-1)},z,u_1^{(t)},u_2^{(t)},v,\delta)$ 抽 $\mu_2^{(t)}$。

（5）从 $\pi(\sigma_1^2|k^{(t-1)},\mu_1^{(t-1)},\mu_2^{(t-1)},\sigma_2^{2(t-1)},z,u_1^{(t)},u_2^{(t)},v,\delta)$ 抽 $\sigma_1^{2(t)}$。

（6）从 $\pi(\sigma_2^2|k^{(t-1)},\mu_1^{(t-1)},\mu_2^{(t-1)},\sigma_1^{2(t-1)},z,u_1^{(t)},u_2^{(t)},v,\delta)$ 抽 $\sigma_2^{2(t)}$。

（7）$k^{(t)}\sim\pi\left(k|,\mu_1^{(t-1)},\mu_2^{(t-1)},\sigma_1^{2(t-1)},\sigma_2^{2(t-1)},z,u_1^{(t)},u_2^{(t)},v,\delta\right)\triangleq\pi\left(k|\bullet\right)$，选分布 $q(k_1^{(t-1)},k_1')$ 取值为 1,2,⋯,$n^{(t-1)}$-1 的离散型均匀分布，即有 $q\left(k^{(t-1)},k_1^{'}\right)=\frac{1}{\left(n^{(t-1)}-1\right)}$，令 $\alpha\left(k^{(t-1)},k^{'}\right)=\min\left\{\frac{\pi\left(k^{'}|\bullet\right)}{\pi\left(k^{(t-1)}|\bullet\right)},1\right\}$，从 1,2,⋯,$n^{(t-1)}$-1 中任意抽取一个 k'，然后由均匀分布 $U(0,1)$ 抽取 u，若 $u\le\alpha(k^{(t-1)},k')$，则 $k^{(t)}=k'$，否则 $k^{(t)}=k^{(t-1)}$。

然后，由以上步骤可得出 $k,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2$ 的一个联合样本 $(k^{(t)},\ \mu_1^{(t)},\mu_2^{(t)},\sigma_1^{2(t)},\sigma_2^{2(t)})$，通过进行 t 次迭代 M 次的 Gibbs 抽样过程，即获取 M 个 5 维独立且同分布的随机样本。

最后，设 $\left(k^{(j)},\mu_1^{(j)},\mu_2^{(j)},\sigma_1^{2(j)},\sigma_2^{2(j)}\right),j=1,2,\cdots,B,\cdots,M$ 为一个容量为 M 的 Gibbs 样本，Gibbs 抽样收敛在第 B 次以后，由于在初始阶段，数据的波动较大，序列不平稳，会对实验的最终结果产生不理想的影响。因此，在计算过程中，需要舍弃前 M 个样本，用 M-B 个样本的均值作各个参数的贝叶斯估计，即：

$$\begin{aligned}&\hat{k}=\frac{1}{M-B}\sum_{j=B+1}^{M}k^{(j)};\\&\hat{\mu_i}=\frac{1}{M-B}\sum_{j=B+1}^{M}\mu_i^{(j)},m=1,2.\\&\hat{\sigma_i^2}=\frac{1}{M-B}\sum_{M-B}^{1}\sigma_i^{2(j)},m=1,2.\end{aligned} \tag{21}$$

4 随机模拟

下面进行随机模拟试验，取试验样本为 n=400。

$$X_i\sim\begin{cases}N(0.6,2) & (i=1,\cdots,150)\\N(5,1.6) & (i=151,\cdots,400)\end{cases} \tag{22}$$

右删失变量 $Y_i\sim N(6,2.5)$，左截断变量 $T_i\sim N(1.5,0.5)$。则 $(k,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)$ 的真实值为 (k=150, μ_1=0.6, μ_2=5, σ_1^2=2, σ_2^2=1.6)。

根据各参数的满条件分布进行 MCMC 模拟，在模拟过程中，丢弃 10 000 次预迭代，迭代过程从 10 001 次到 20 000 次运行结果如表 1 所示。

表 1 参数 $k,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2$ 的 bayes 估计

参数	真值	均值	MC 误差	相对误差	97.5% 分位数	中位数	2.5% 分位数
k	150	150.912 5	0.041 2	0.006	152	150	147
μ_1	0.8	0.810 2	0.003 6	0.012 7	0.851 7	0.808 5	0.770 4
μ_2	0.9	0.944 1	0.003 1	0.049	0.983 1	0.943 6	0.840 1
σ_1^2	4	3.859 3	0.003 5	0.035 1	4.109 2	3.861 8	3.587 2
σ_2^2	1	0.992 7	0.001 4	0.007 3	1.213 0	0.993 1	0.931 6

在模型分析中 MCMC 的收敛性诊断非常重要。检查多个独立的 MCMC 链是否收敛到相同的稳定分布，在收敛性诊断中，得到缩放因子诊断值为 0.975，结果接近于 1，这表示着 MCMC 算法已经收敛良好。不同链之间的变异性不显著高于链内变异性，所得结果可靠，结果如图 1 所示。

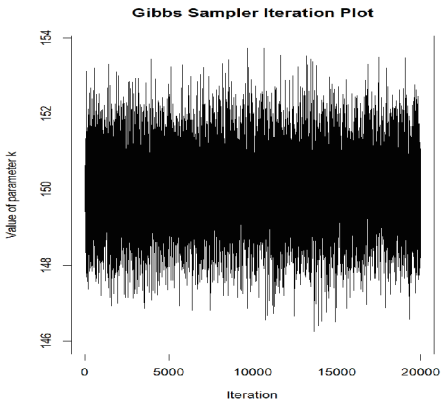

图 1 参数 k 的 Gibbs 迭代抽样过程图

5 实证分析

根据 sklean 数据集[19]中关于对糖尿病患者进行血糖水平记录的前 100 个数据进行研究，在记录的过程中，可能因为临床决策因素、数据采集和记录的偏向性导致删除了某些数据，利用左截断右删失数据下正态分布参数单变点的贝叶斯估计方法对现有信息探索数据的规律。

5.1 正态性检验

对糖尿病患者的血糖水平数据记为 Blood Sugar，采用 Shapior-Wilk 检验对所取数据进行正态性检验，假设检验如下。

H_0：此组数据服从正态分布。

H_1：此组数据不服从正态分布。

利用 Shapior-Wilk 检验方法得到结果如表 2 所示。

表 2 Shapior-Wilk normality test

数据	P-value
Blood Sugar	0.425 1

由表 2 所得，Shapior-Wilk 检验的 P 值大于 0.05，说明原假设成立，则这组数据服从正态分布。因此，可以利用此数据对本文所提出的正态分布的未知参数进行估计。

5.2 模型参数的估计

由实证所用数据所得，在模拟过程中，从第 10 001 次开始到第 20 000 次的运行结果如表 3 所示。

表 3 参数 $k_1, \mu_3, \mu_4, \sigma_3^2, \sigma_4^2$ 的 bayes 估计

参数	真值	均值	MC 误差	相对误差	97.5% 分位数	中位数	2.5% 分位数
k_1	78	78.612 5	0.001 4	0.007 8	79	78	76
μ_3	89	86.314 3	0.002 2	0.030 1	87.117 9	86.274 8	85.541 7
μ_4	85	83.163 4	0.003 7	0.021 3	84.133 6	83.201 9	82.283 7
σ_3^2	106	102.665 7	0.004 3	0.031 5	103.805 7	102.638 4	101.471 8
σ_4^2	163	154.693 8	0.004 6	0.050 9	156.348 1	154.687 5	153.147 9

对实证结果进行收敛性诊断，得到缩放因子诊断值为 0.928，说明收敛效果良好，实证数据参数的真实值与参数的贝叶斯估计值的模拟精度较高，结果如图 2 所示。

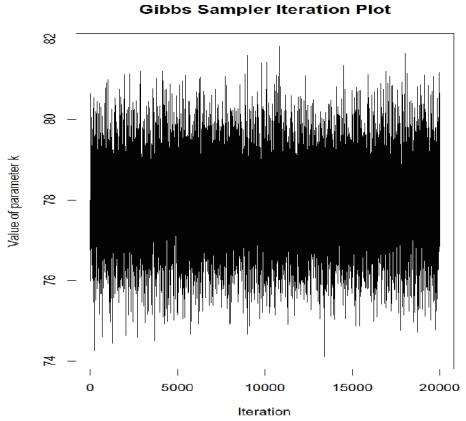

图 2 参数 k_1 的 Gibbs 抽样迭代过程图

本文研究的应用潜力在于能够在考虑数据缺失的情况下，准确估计正态分布的参数，为实际生活中的寿命实验和其他领域的数据分析提供有力工具。

参考文献：

[1] PAGE E S. Continuous inspection schemes[J]. Biometrika, 1954, 41(12): 100-115.

[2] CSORGO M，HORVÁTH L.Limit Theorems in Change-Point Analysis[EB/OL].(1997-11-29)[2023-09-26].https://www.semanticscholar.org/paper/Limit-Theorems-in-Change-Point-Analysis-Csorgo-Horv%C3%A1th/71421b7b6ebccebb74420b1c940134c7ff3cdb43.

[3] LIANG F, WONG W H. Real: parameter evolutionary Monte Carlo with applications to Bayesian mixture models[J]. Journal of the American Statistical Association, 2001, 96(454): 653-666.

[4] CHERNOF H, ZACKS S. Estimating the current mean of a normal distribution which is subjected to changes in time[J]. The annals of mathematical statistics,1964,35(3):999-1018.

[5] PERREAUH L,BERNIER J,BOBRE B,et a1. Bayesian change:point analysis in hydrometeorological time series(Part 1)：the normal model revisited[J]. Journal of hydrology, 2000, 235(3):221-241.

[6] HINKLEY D V．Inference about the change: point in a sequence of random variables[J]. Biometrika, 1970,57(1): 1-17.

[7] SHIMIZU R. A characterization of the normal distribution[J]. Ann Inst Stat Math,1961,13:53-56.

[8] Gupta B N. A characterisation of the normal distribution[J]. Trab. Estad. Invest. Oper. 1970, 21: 35-38.

[9] ANSCOMBE F J. Normal likelihood functions[J]. Ann Inst Stat Math, 1964,16:1-19 .

[10] BALAKRISHNAN N,MITRA D．Likelihood inference for lognormal data with left truncation and right censoring with an illustration[J]. Journal of statistical planning and inference, 2011,141(11): 3536-3553.

[11] GROSS S T, LAI T L．Nonparametric estimation and regression analysis with left—truncated and right—censored data[J]. Journal of the American Statistical Association, 1996, 91(435): 1166-1180.

基于 MPI 的 OTP 三角形计数算法研究

龙昌庭[1]
LONG Changting

摘 要 在社交网络分析、推荐系统和聚类系数等大规模图分析问题中，计算图中三角形的数量是一项重要的任务。然而，当面临大量数据以及子图之间存在重复的三角形结构时，计算变得困难且具有挑战性。因此，图数据的分布式计算变成了研究热点。提出一种基于 OTP 三角形计数算法的 MPI 优化算法，OTP 算法是基于 MapReduce 框架的三角形计数算法，但在三角形数量的计算过程中，计算时间仍然过长。通过实验结果的分析，发现优化后的算法相较于现有算法，在计算时间上显著缩短 10 ～ 40 倍，特别是在处理非常大规模图时。这一优化进一步弥补了现有算法在执行时间性能方面的不足。

关键词 图划分；三角形计数；分布式计算；邻接表；MapReduce；消息传递

doi：10.3969/j.issn.1672-9528.2024.02.029

0 引言

近年来，随着互联网信息技术的迅速发展，各行各业产生了大量的数据。图这种特殊的数据结构，不仅可以存储数据本身，还可以存储数据之间的复杂关联关系，因此受到了学术界和工业界的广泛关注。图可以表示许多重要的数据，如社交网络、生物信息网络、交通网络等，分析图数据可以得到重要的洞见。学术界研究图挖掘算法，以发现图数据中的模式和规律。工业界利用图分析技术，获得重要的商业智能，改进决策。总之，图作为一种关键的数据结构，能表示现实世界的复杂关系，分析图数据可以获得重要知识，因此研究图数据具有重要意义，受到学术界和工业界的广泛重视。

三角形数量是图分析中的一个重要任务，它主要用于识别和计算图中存在的三角形结构。通过对三角形数量的研究，可以度量网络的聚类系数，即节点之间的紧密程度。聚类系数可以帮助理解网络的群体结构、信息传播和社交影响力等，三角形数量是研究和分析复杂网络的重要指标[1]。在异常检测[2]中，也可以通过计算图中的三角形数量和分布情况，可以发现与正常情况不符的异常模式，在推荐系统[3]中，通过

1. 贵州财经大学信息学院 贵州贵阳 550025

[12] MOLANES L E M, CAO R, KEILEGOM I V. Smoothed empirical likelihood confidence intervals for the relative distribution with left-truncated and fight-censored data[J]. Canadian journal of statistics, 2010,38(3): 453-473.

[13] 茆诗松 , 汤银才 . 贝叶斯统计 [M]. 北京 : 中国统计出版社 , 2012.

[14] 何朝兵 , 刘华文 . 左截断右删失数据下对数正态分布参数多变点的贝叶斯估计 [J]. 福州大学学报 (自然科学版), 2014, 42(4):507-513.

[15] 彭秋曦 . 左截断右删失数据下指数分布变点的 Bayes 估计 [D]. 重庆 : 重庆大学 ,2015.

[16] 何朝兵 , 刘华文 . 左截断右删失数据下二项分布参数多变点的贝叶斯估计 [J]. 华南师范大学学报 (自然科学版), 2014, 46(3):34-38.

[17] 何朝兵 , 刘华文 . 左截断右删失数据下几何分布参数多变点的贝叶斯估计 [J]. 重庆师范大学学报 (自然科学版), 2014, 31(4):100-105.

[18] 梅梦玲 .IIRCT 下某些指数型分布族参数变点的贝叶斯估计 [D]. 乌鲁木齐 : 新疆师范大学 ,2021.

[19] BRADLEY E, TREVOR H I, ROBERT T. Least Angle Regression[EB/OL]. (2004-06-23)[2023-10-06]. https://arxiv.org/abs/math/0406456.

【作者简介】

李清宇（1999—），女，贵州铜仁人，硕士，研究方向：统计模型与统计计算。

黄介武（1977—），通信作者（email: 846221886@qq.com），男，湖南沅江人，博士，教授，研究方向：统计模型与统计计算。

（收稿日期：2023-11-27）