应用概率统计　第二十八卷
第六期　2012年12月

Chinese Journal of Applied Probability
and Statistics Vol.28 No.6 Dec. 2012



# 数据分组和右删失下混合广义指数分布的参数估计 *

田玉柱[1,2]　田茂再[2]　陈　平[3]

([1]天水师范学院数学与统计学院, 天水, 741001; [2]中国人民大学统计学院, 北京, 100872)

([3]东南大学数学系, 南京, 210096)

**摘　要**

本文利用EM算法考虑了混合广义指数分布模型在分组数据和右删失情形下的参数估计问题, 并给出了相应的估计公式, 一组模拟研究说明了所提算法的有效性. 最后, 应用所提模型对一组医学数据进行了分析.

**关键词:** 混合广义指数分布, 分组数据和右删失数据, EM算法.

**学科分类号:** O213.

## §1. 引　　言

Gupta和Kundu (1999)提出了双参数广义指数分布并把它作为Gamma分布和Weibull分布的替代分布. 广义指数分布在寿命数据分析和可靠性分析等领域有广泛的应用. 单总体广义指数分布的研究已有许多, 如Raqab (2002), Sarhan (2007), Gupta和Kundu (2007), Gupta和Kundu (2008), Kundu和Pradhan (2009), Chen和Lio (2010)等, 但实际问题中混合模型更为灵活. 混合模型已被广泛应用到产品寿命分析、临床医学、图像处理等领域. 一般混合模型及混合广义指数分布的讨论已有许多结果, 如Sultan, Ismail和Al-Moisheer(2007), Jones和Ashour(1976), Elsherpieny(2007)等. 在寿命数据和医学研究中, 经常会碰到数据分组或删失的情形. 单分布模型在数据分组和右删失下的估计方法已有一定研究, 见Pettitt (1985), Liu (2001), Liu, Chen和Fei (2008)等, 而混合模型在数据分组和右删失下的研究还相当罕见. 因此, 本文将考虑数据分组和右删失情形下混合广义指数分布的参数估计问题, 模型的概率密度函数和分布函数分别是

$$f(x;p,\alpha,\lambda)=\sum_{k=1}^{m}p_k\alpha_k\lambda_k(1-e^{-\lambda_k x})^{\alpha_k-1}e^{-\lambda_k x},\qquad x\geq 0; \tag{1.1}$$

$$F(x;p,\alpha,\lambda)=\sum_{k=1}^{m}p_k(1-e^{-\lambda_k x})^{\alpha_k},\qquad x\geq 0, \tag{1.2}$$

其中$p=(p_1,\ldots,p_{m-1})$, $\alpha=(\alpha_1,\ldots,\alpha_m)$, $\lambda=(\lambda_1,\ldots,\lambda_m)$, 并且$0<p_k<1$, $k=1,\ldots,m-1$, $p_m=1-\sum_{k=1}^{m-1}p_k$; $\alpha_k>0$, $\lambda_k>0$, $k=1,2,\ldots,m$, 共有$3m-1$个参数. 本文第2节简单介绍了参数的极大似然估计. 第3节利用EM算法考虑了模型的参数估计. 第4节通过数值模拟说明了算法的有效性. 第5节对一组医学数据进行了分析.

*教育部科学技术研究重点项目(108120)和中国人民大学研究生科学研究基金项目(12XNH161)资助.
本文2010年9月24日收到, 2012年8月11日收到修改稿.

## §2. 数据分组和右删失情形下模型的对数似然函数

取$n$个产品进行寿命检验, 获得数据如下: 将$(0,+\infty)$分成$N+1$个区间, 前$N$个区间记为$(T_{i-1},T_i]$, 其中$i=1,2,\ldots,N$; $0=T_0<T_1<\cdots<T_N<+\infty$. 再用$c_i$表示落入到$(T_{i-1},T_i]$中的失效产品数, 用$d_i$表示在区间$(T_{i-1},T_i]$的右端点$T_i$被删失的产品个数. 显然有$n=\sum\limits_{i=1}^{N}(c_i+d_i)$.

模型的似然函数为

$$\begin{aligned} L(p,\alpha,\lambda) &= \prod_{i=1}^{N}[F(T_i;p,\lambda)-F(T_{i-1};p,\lambda)]^{c_i}\cdot[1-F(T_i;p,\lambda)]^{d_i} \\ &= \prod_{i=1}^{N}\Big[\sum_{k=1}^{m}p_k((1-e^{-\lambda_k T_i})^{\alpha_k}-(1-e^{-\lambda_k T_{i-1}})^{\alpha_k})\Big]^{c_i} \\ &\quad\cdot\Big[1-\sum_{k=1}^{m}p_k(1-e^{-\lambda_k T_i})^{\alpha_k}\Big]^{d_i}. \end{aligned}$$

对数似然函数为

$$\begin{aligned} \log L(p,\alpha,\lambda) &= \sum_{i=1}^{N}\Big\{c_i\cdot\log\Big[\sum_{k=1}^{m}p_k((1-e^{-\lambda_k T_i})^{\alpha_k}-(1-e^{-\lambda_k T_{i-1}})^{\alpha_k})\Big] \\ &\quad+d_i\cdot\log\Big[1-\sum_{k=1}^{m}p_k(1-e^{-\lambda_k T_i})^{\alpha_k}\Big]\Big\}. \end{aligned}$$

注意到上述对数似然函数相当复杂, 直接求导数无法得到估计的显式表达, 即使运用数值算法也想当复杂. 下面利用EM算法更方便地求解参数的极大似然估计(MLE).

## §3. 基于EM算法的参数估计

EM算法是由Dempster等人于1977提出的一种用来求解有缺失数据(missing data), 删失数据以及带有噪声等所谓的不完全数据模型MLE的迭代算法, 它主要利用已获得的观测数据(observed data)来求MLE. EM算法是一种迭代算法, 每一次迭代都能保证似然函数值增加, 并且收敛到一个局部极大值. EM算法包括两步: 第一步求期望(Expectation Step), 称为E步; 第二步求极大值(Maximization Step), 称为M步. 关于EM算法的求解步骤可参见Dempster (1977)等.

设某$n$个产品的寿命$X_1,X_2,\ldots,X_n$独立同分布于混合广义指数分布模型(1.1), 对于随机变量$X_j$, 我们记

$$\begin{aligned} &f_{kj}=\alpha_k\lambda_k(1-e^{-\lambda_k x_j})^{\alpha_k-1}e^{-\lambda_k x_j}, \quad s_{kj}=1-(1-e^{-\lambda_k x_j})^{\alpha_k}, \qquad k=1,\ldots,m; \\ &f_j=\sum_{k=1}^{m}p_k f_{kj}, \quad s_j=\sum_{k=1}^{m}p_k s_{kj}, \qquad j=1,\ldots,n, \end{aligned}$$

其中$s_x$表示产品的生存函数.

再引入随机变量$X_j$的示性向量$I_j = (I_{j1}, \ldots, I_{jm})$, 其中$I_j$的分量中仅有一个为1, 其余均为0, 且若$I_{jk} = 1$, 即说明$X_j$来自混合分布中的第$k$个子总体. 记$I = (I_1, \ldots, I_n)$, 则$I$可表示所有产品的寿命$X_1, X_2, \ldots, X_n$的示性向量. 对于$X_j$的示性向量$I_j$而言, 显然$I_j = (I_{j1}, \ldots, I_{jm})$服从多点分布, 但是寿命观测$X_j$来自哪个子总体并不知道, 即$I_j$是不可观测的, 那么$I = (I_1, \ldots, I_n)$也是不可观测的, 它在EM算法中被视为缺失数据. 为了下文讨论的方便, 我们用$I_j^{(1)} = (I_{j1}^{(1)}, \ldots, I_{jm}^{(1)})$和$I_j^{(2)} = (I_{j1}^{(2)}, \ldots, I_{jm}^{(2)})$分别表示完全寿命数据和右删失数据的示性向量.

对于完全寿命数据, $X_j$和$I_j^{(1)}$的联合分布是

$$g(x_j, I_j^{(1)}|p, \alpha, \lambda) = \prod_{k=1}^{m} [p_k f_{kj}]^{I_{jk}^{(1)}},$$

则$I_j^{(1)}$在$X_j$给定时的条件分布是

$$P(I_{jk}^{(1)} = 1|x_j, p, \alpha, \lambda) = \frac{p_k f_{kj}}{f_j}, \qquad k = 1, 2, \ldots, m.$$

对于右删失数据, $X_j$和$I_j^{(2)}$的联合分布是

$$g(x_j, I_j^{(2)}|p, \alpha, \lambda) = \prod_{k=1}^{m} [p_k s_{kj}]^{I_{jk}^{(2)}},$$

则$I_j^{(2)}$在$X_j$给定时的条件分布是

$$P(I_{jk}^{(2)} = 1|x_j, p, \alpha, \lambda) = \frac{p_k s_{kj}}{s_j}, \qquad k = 1, 2, \ldots, m.$$

将这$n$个产品进行寿命试验, 它们分别落入区间$(T_{i-1}, T_i]$或在$T_i$时刻被删失, 我们只能观测到产品寿命落入区间$(T_{i-1}, T_i]$中$X_j$的个数$c_i$以及在$T_i$时刻被删失$X_j$的个数$d_i$, 其中$i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, n$; $0 = T_0 < T_1 < \cdots < T_N < +\infty$. 记产品的寿命全体为$X = (X_1, X_2, \ldots, X_n)$, 寿命试验中$X$也是不可观测的, 能观测到仅为$Y = (c_1, \ldots, c_N, d_1, \ldots, d_N)$, 此时缺失数据为$(X, I)$, 这样完全数据可表示$Z = (X, I, Y)$. 为了应用EM算法, 我们再引入随机变量$X_{ih}, X_{il}$, 它们分别表示落入区间$(T_{i-1}, T_i]$和在$T_i$时刻被删失的产品寿命. 下面我们根据EM算法中的E步和M步来获得被估参数的极大似然估计.

事实上, $X$已经包含了观测结果$Y$所有的信息, 于是完全数据的似然函数可记为$f(p, \alpha, \lambda|X, I, Y) = f(p, \alpha, \lambda|X, I)$, 即

$$\begin{aligned} f(p, \alpha, \lambda|X, I) &= \prod_{i=1}^{N} \Big\{ \Big[ \prod_{k=1}^{m} (p_k \alpha_k \lambda_k (1 - e^{-\lambda_k x_{ih}})^{\alpha_k - 1} e^{-\lambda_k x_{ih}})^{I_{ik}^{(1)}} \Big]^{c_i} \\ &\quad \cdot \Big[ \prod_{k=1}^{m} (p_k \alpha_k \lambda_k (1 - e^{-\lambda_k x_{il}})^{\alpha_k - 1} e^{-\lambda_k x_{il}})^{I_{ik}^{(2)}} \Big]^{d_i} \Big\} \\ &= \prod_{i=1}^{N} \prod_{k=1}^{m} [(p_k \alpha_k \lambda_k (1 - e^{-\lambda_k x_{ih}})^{\alpha_k - 1} e^{-\lambda_k x_{ih}})^{I_{ik}^{(1)} c_i} \\ &\quad \cdot (p_k \alpha_k \lambda_k (1 - e^{-\lambda_k x_{il}})^{\alpha_k - 1} e^{-\lambda_k x_{il}})^{I_{ik}^{(2)} d_i}]. \end{aligned}$$

则完全数据的对数似然是

$$\begin{aligned}\log f(p,\alpha,\lambda|X,I) &= \sum_{i=1}^{N}\sum_{k=1}^{m}[c_iI_{ik}^{(1)}(\log(p_k\lambda_k\alpha_k)-\lambda_kx_{ih}+(\alpha_k-1)\cdot\log(1-e^{-\lambda_kx_{ih}}))\\&\quad+d_iI_{ik}^{(2)}(\log(p_k\lambda_k\alpha_k)-\lambda_kx_{il}+(\alpha_k-1)\cdot\log(1-e^{-\lambda_kx_{il}}))].\end{aligned}$$

设定初值$p^{(0)},\alpha^{(0)},\lambda^{(0)}$, EM算法的步骤为:

E步: 假定参数的第$t-1$步估计$p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)}$, 则第$t$步的$Q$函数为

$$\begin{aligned}&Q(p,\alpha,\lambda|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\=\;&\mathsf{E}[\log f(p,\alpha,\lambda|X,I)|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y]\\=\;&\sum_{i=1}^{N}\sum_{k=1}^{m}\mathsf{E}\{[c_iI_{ik}^{(1)}(\log(p_k\lambda_k\alpha_k)-\lambda_kx_{ih}+(\alpha_k-1)\cdot\log(1-e^{-\lambda_kx_{ih}}))\\&+d_iI_{ik}^{(2)}(\log(p_k\lambda_k\alpha_k)-\lambda_kx_{il}+(\alpha_k-1)\cdot\log(1-e^{-\lambda_kx_{il}}))]|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y\}\\=\;&\sum_{i=1}^{N}\sum_{k=1}^{m}[c_i\log(p_k\lambda_k\alpha_k)\cdot\mathsf{E}(I_{ik}^{(1)}|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\&-c_i\lambda_k\cdot\mathsf{E}(I_{ik}^{(1)}x_{ih}|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\&+c_i(\alpha_k-1)\cdot\mathsf{E}(I_{ik}^{(1)}\log(1-e^{-\lambda_kx_{ih}})|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\&+d_i\log(p_k\lambda_k\alpha_k)\cdot\mathsf{E}(I_{ik}^{(2)}|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\&-d_i\lambda_k\cdot\mathsf{E}(I_{ik}^{(2)}x_{il}|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\&+d_i(\alpha_k-1)\cdot\mathsf{E}(I_{ik}^{(2)}\log(1-e^{-\lambda_kx_{il}})|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\=\;&\sum_{i=1}^{N}\sum_{k=1}^{m}[c_i\log(p_k\lambda_k\alpha_k)\cdot\mathsf{E}(\mathsf{E}(I_{ik}^{(1)}|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},X))\\&-c_i\lambda_k\cdot\mathsf{E}(\mathsf{E}(I_{ik}^{(1)}x_{ih}|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},X))\\&+c_i\cdot(\alpha_k-1)\mathsf{E}(\mathsf{E}(I_{ik}^{(1)}\log(1-e^{-\lambda_kx_{ih}})|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},X))\\&+d_i\log(p_k\lambda_k\alpha_k)\cdot\mathsf{E}(\mathsf{E}(I_{ik}^{(2)}|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},X))\\&-d_i\lambda_k\cdot\mathsf{E}(\mathsf{E}(I_{ik}^{(2)}x_{il}|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},X))\\&+d_i(\alpha_k-1)\cdot\mathsf{E}(\mathsf{E}(I_{ik}^{(2)}\log(1-e^{-\lambda_kx_{ih}})|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},X))]\\=\;&\sum_{i=1}^{N}\sum_{k=1}^{m}[c_i\log(p_k\lambda_k\alpha_k)\cdot\mathsf{E}(a_{ki}^{(t-1)}(x_{ih}))-c_i\lambda_k\cdot\mathsf{E}(x_{ih}\cdot a_{ki}^{(t-1)}(x_{ih}))\\&+c_i(\alpha_k-1)\cdot\mathsf{E}(\log(1-e^{-\lambda_kx_{ih}})\cdot a_{ki}^{(t-1)}(x_{ih}))+d_i\log(p_k\lambda_k\alpha_k)\cdot\mathsf{E}(b_{ki}^{(t-1)}(x_{il}))\\&-d_i\lambda_k\cdot\mathsf{E}(x_{il}\cdot b_{ki}^{(t-1)}(x_{il}))+d_i(\alpha_k-1)\cdot\mathsf{E}(\log(1-e^{-\lambda_kx_{ih}})\cdot b_{ki}^{(t-1)}(x_{il}))],\end{aligned}$$

其中

$$f_{ki}^{(t-1)}=\alpha_k^{(t-1)}\lambda_k^{(t-1)}(1-e^{-\lambda_k^{(t-1)}x_{ih}})^{\alpha_k^{(t-1)}-1}e^{-\lambda_k^{(t-1)}x_{ih}},\qquad f_i^{(t-1)}=\sum_{k=1}^{m}p_k^{(t-1)}f_{ki}^{(t-1)},$$

$$s_{ki}^{(t-1)}=1-(1-e^{-\lambda_k^{(t-1)}x_{il}})^{\alpha_k^{(t-1)}},\qquad s_i^{(t-1)}=\sum_{k=1}^{m}p_k^{(t-1)}s_{ki}^{(t-1)},$$

$$a_{ki}^{(t-1)}(x_{ih})=\frac{p_k^{(t-1)}f_{ki}^{(t-1)}}{f_i^{(t-1)}},\quad b_{ki}^{(t-1)}(x_{il})=\frac{p_k^{(t-1)}s_{ki}^{(t-1)}}{s_i^{(t-1)}};\quad k=1,2,\ldots,m,\ i=1,2,\ldots,N.$$

在上述$Q$函数中, $X_{ih}$和$X_{il}$的条件概率密度函数分别记为$p_{ih}(x)$和$p_{il}(x)$, 即

$$\begin{aligned}
p_{ih}(x) &= f_{ih}(x|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\
&= \frac{\sum_{k=1}^{m}p_k^{(t-1)}\alpha_k^{(t-1)}\lambda_k^{(t-1)}(1-e^{-\lambda_k^{(t-1)}x})^{\alpha_k^{(t-1)}-1}e^{-\lambda_k^{(t-1)}x}}{\sum_{k=1}^{m}p_k^{(t-1)}((1-e^{-\lambda_k^{(t-1)}T_i})^{\alpha_k^{(t-1)}}-(1-e^{-\lambda_k^{(t-1)}T_{i-1}})^{\alpha_k^{(t-1)}})},\qquad x\in(T_{i-1},T_i],\\
p_{il}(x) &= f_{il}(x|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\
&= \frac{\sum_{k=1}^{m}p_k^{(t-1)}\alpha_k^{(t-1)}\lambda_k^{(t-1)}(1-e^{-\lambda_k^{(t-1)}x})^{\alpha_k^{(t-1)}-1}e^{-\lambda_k^{(t-1)}x}}{1-\sum_{k=1}^{m}p_k^{(t-1)}(1-e^{-\lambda_k^{(t-1)}T_i})^{\alpha_k^{(t-1)}}},\qquad x\in(T_i,+\infty).
\end{aligned}$$

于是得

$$\begin{aligned}
&Q(p,\alpha,\lambda|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)\\
=&\sum_{i=1}^{N}\sum_{k=1}^{m}\Big[c_i\log(p_k\lambda_k\alpha_k)\cdot\int_{T_{i-1}}^{T_i}a_{ki}^{(t-1)}(x)\cdot p_{ih}(x)\mathrm{d}x-c_i\lambda_k\cdot\int_{T_{i-1}}^{T_i}x\cdot a_{ki}^{(t-1)}(x)\cdot p_{ih}(x)\mathrm{d}x\\
&+c_i(\alpha_k-1)\cdot\int_{T_{i-1}}^{T_i}a_{ki}^{(t-1)}(x)\cdot\log(1-e^{-\lambda_k x})\cdot p_{ih}(x)\mathrm{d}x\\
&+d_i\log(p_k\lambda_k\alpha_k)\cdot\int_{T_i}^{\infty}b_{ki}^{(t-1)}(x)\cdot p_{il}(x)\mathrm{d}x-d_i\lambda_k\cdot\int_{T_i}^{\infty}b_{ki}^{(t-1)}(x)\cdot x\cdot p_{il}(x)\mathrm{d}x\\
&+d_i(\alpha_k-1)\cdot\int_{T_i}^{\infty}b_{ki}^{(t-1)}(x)\cdot\log(1-e^{-\lambda_k x})\cdot p_{il}(x)\mathrm{d}x\Big]\\
=&\sum_{i=1}^{N}\sum_{k=1}^{m}[\log(p_k\lambda_k\alpha_k)\cdot(c_i\cdot\triangle1_{ki}^{(t-1)}+d_i\cdot\triangle4_{ki}^{(t-1)})-\lambda_k\cdot(c_i\cdot\triangle2_{ki}^{(t-1)}+d_i\cdot\triangle5_{ki}^{(t-1)})\\
&+(\alpha_k-1)\cdot(c_i\cdot\triangle3_{ki}^{(t-1)}+d_i\cdot\triangle6_{ki}^{(t-1)})],
\end{aligned}$$

其中

$$\triangle1_{ki}^{(t-1)}=\int_{T_{i-1}}^{T_i}a_{ki}^{(t-1)}(x)\cdot p_{ih}(x)\mathrm{d}x,\qquad\triangle2_{ki}^{(t-1)}=\int_{T_{i-1}}^{T_i}x\cdot a_{ki}^{(t-1)}(x)\cdot p_{ih}(x)\mathrm{d}x,$$

$$\triangle3_{ki}^{(t-1)}=\int_{T_{i-1}}^{T_i}\log(1-e^{-\lambda_k x})\cdot a_{ki}^{(t-1)}(x)\cdot p_{ih}(x)\mathrm{d}x,$$

$$\triangle4_{ki}^{(t-1)}=\int_{T_i}^{+\infty}b_{ki}^{(t-1)}(x)\cdot p_{il}(x)\mathrm{d}x,\qquad\triangle5_{ki}^{(t-1)}=\int_{T_i}^{+\infty}x\cdot b_{ki}^{(t-1)}(x)\cdot p_{il}(x)\mathrm{d}x,$$

$$\triangle6_{ki}^{(t-1)}=\int_{T_i}^{+\infty}\log(1-e^{-\lambda_k x})\cdot b_{ki}^{(t-1)}(x)\cdot p_{il}(x)\mathrm{d}x.$$

M步: 极大化$Q$函数得参数$p,\alpha,\lambda$的第$t$步估计$p^{(t)},\alpha^{(t)},\lambda^{(t)}$, 即将$Q(p,\alpha,\lambda|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)$分别对参数$p,\alpha,\lambda$求导后得$Q(p,\alpha,\lambda|p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)},Y)$的极大值点$p^{(t)},\alpha^{(t)},\lambda^{(t)}$.

为了数值计算的方便, 我们把上述$Q$函数中右端式子中的

$$\begin{aligned}\triangle 3_{ki}^{(t-1)}&=\int_{T_{i-1}}^{T_i}\log(1-e^{-\lambda_k x})\cdot a_{ki}^{(t-1)}(x)\cdot p_{ih}(x)\mathrm{d}x,\\ \triangle 6_{ki}^{(t-1)}&=\int_{T_i}^{+\infty}\log(1-e^{-\lambda_k x})\cdot b_{ki}^{(t-1)}(x)\cdot p_{il}(x)\mathrm{d}x\end{aligned}$$

替换为

$$\begin{aligned}\widetilde{\triangle} 3_{ki}^{(t-1)}&=\int_{T_{i-1}}^{T_i}\log(1-e^{-\lambda_k^{(t-1)} x})\cdot a_{ki}^{(t-1)}(x)\cdot p_{ih}(x)\mathrm{d}x,\\ \widetilde{\triangle} 6_{ki}^{(t-1)}&=\int_{T_i}^{+\infty}\log(1-e^{-\lambda_k^{(t-1)} x})\cdot b_{ki}^{(t-1)}(x)\cdot p_{il}(x)\mathrm{d}x.\end{aligned}$$

然后对$\alpha,\lambda,p$分别求导, 并令

$$\begin{aligned}\frac{\partial Q}{\partial\alpha_k}&=\sum_{i=1}^{N}\Big[\frac{1}{\alpha_k}(c_i\cdot\triangle 1_{ki}^{(t-1)}+d_i\cdot\triangle 4_{ki}^{(t-1)})+(c_i\cdot\widetilde{\triangle} 3_{ki}^{(t-1)}+d_i\cdot\widetilde{\triangle} 6_{ki}^{(t-1)})\Big]\\ &=0,\qquad k=1,\ldots,m;\end{aligned}\tag{3.1}$$

$$\begin{aligned}\frac{\partial Q}{\partial\lambda_k}&=\sum_{i=1}^{N}\Big[\frac{1}{\lambda_k}(c_i\cdot\triangle 1_{ki}^{(t-1)}+d_i\cdot\triangle 4_{ki}^{(t-1)})-(c_i\cdot\triangle 2_{ki}^{(t-1)}+d_i\cdot\triangle 5_{ki}^{(t-1)})\Big]\\ &=0,\qquad k=1,\ldots,m;\end{aligned}\tag{3.2}$$

$$\begin{aligned}\frac{\partial Q}{\partial p_k}&=\sum_{i=1}^{N}\left[\frac{1}{p_k}(c_i\cdot\triangle 1_{ki}^{(t-1)}+d_i\cdot\triangle 4_{ki}^{(t-1)})-\frac{1}{1-\sum_{j=1}^{m-1}p_j}(c_i\cdot\triangle 1_{mi}^{(t-1)}+d_i\cdot\triangle 4_{mi}^{(t-1)})\right]\\ &=0,\qquad k=1,\ldots,m-1.\end{aligned}\tag{3.3}$$

由(3.1)可得

$$\alpha_k^{(t)}=-\frac{\sum_{i=1}^{N}[c_i\cdot\triangle 1_{ki}^{(t-1)}+d_i\cdot\triangle 4_{ki}^{(t-1)})}{\sum_{i=1}^{N}(c_i\cdot\widetilde{\triangle} 3_{ki}^{(t-1)}+d_i\cdot\widetilde{\triangle} 6_{ki}^{(t-1)})},\qquad k=1,\ldots,m.\tag{3.4}$$

由(3.2)可得

$$\lambda_k^{(t)}=\frac{\sum_{i=1}^{N}(c_i\cdot\triangle 1_{ki}^{(t-1)}+d_i\cdot\triangle 4_{ki}^{(t-1)})}{\sum_{i=1}^{N}(c_i\cdot\triangle 2_{ki}^{(t-1)}+d_i\cdot\triangle 5_{ki}^{(t-1)})},\qquad k=1,\ldots,m.\tag{3.5}$$

由(3.3)可得方程组

$$\begin{aligned} & p_k\sum_{i=1}^{N}(c_i\cdot\triangle 1_{mi}^{(t-1)}+d_i\cdot\triangle 4_{mi}^{(t-1)})+\sum_{j=1}^{m-1}p_j\sum_{i=1}^{N}(c_i\cdot\triangle 1_{ki}^{(t-1)}+d_i\cdot\triangle 4_{ki}^{(t-1)}) \\ = & \sum_{i=1}^{N}(c_i\cdot\triangle 1_{ki}^{(t-1)}+d_i\cdot\triangle 4_{ki}^{(t-1)}), \qquad k=1,\ldots,m-1. \end{aligned}$$

整理可得参数$p_1,\ldots,p_{m-1}$的解是非齐次线性方程组$Ap=b$的解, $p,A,b$分别是

$$p=(p_1,p_2,\ldots,p_{m-1})^T, \qquad A_{m-1}=(a_{ls}),$$

$$a_{ls}=\begin{cases}\sum_{i=1}^{N}(c_i\cdot\triangle 1_{li}^{(t-1)}+d_i\cdot\triangle 4_{li}^{(t-1)})+\sum_{i=1}^{N}(c_i\cdot\triangle 1_{mi}^{(t-1)}+d_i\cdot\triangle 4_{mi}^{(t-1)}), & l=s;\\ \sum_{i=1}^{N}(c_i\cdot\triangle 1_{li}^{(t-1)}+d_i\cdot\triangle 4_{li}^{(t-1)}), & l\neq s,\end{cases}$$

$$b=\Big(\sum_{i=1}^{N}(c_i\cdot\triangle 1_{1i}^{(t-1)}+d_i\cdot\triangle 4_{1i}^{(t-1)}),\ldots,\sum_{i=1}^{N}(c_i\cdot\triangle 1_{m-1,i}^{(t-1)}+d_i\cdot\triangle 4_{m-1,i}^{(t-1)})\Big)^T.$$

由于$\sum_{i=1}^{N}(c_i\cdot\triangle 1_{li}^{(t-1)}+d_i\cdot\triangle 4_{li}^{(t-1)})>0$, $l=1,\ldots,m$, 故易证$\mathrm{rank}(A)=m-1$, 即$A$可逆, 则参数向量$p$的第$t$次迭代值为

$$p^{(t)}=(p_1^{(t)},p_2^{(t)},\ldots,p_{m-1}^{(t)})^T=A^{-1}b. \tag{3.6}$$

解出(3.4)(3.5)(3.6)就是所求的$(p^{(t)},\alpha^{(t)},\lambda^{(t)})$, 这样就完成了一次迭代$(p^{(t-1)},\alpha^{(t-1)},\lambda^{(t-1)})\to(p^{(t)},\alpha^{(t)},\lambda^{(t)})$, 不断重复E步和M步, 即重复上面(3.4)(3.5)(3.6)式直至$p,\alpha,\lambda$收敛为止.

## §4. 数值模拟

设$X_i$, $i=1,2,\ldots,n$是来自模型(1.1)的独立同分布样本, 考虑2成分混合即$m=2$时模型(1.1)在分组数据和右删失数据下的参数估计, 为了验证所提算法的正确性, 模型真参数假定为$p_1=0.4$, $\alpha_1=0.8$, $\alpha_2=0.5$, $\lambda_1=0.06$, $\lambda_2=0.004$, 将样本数据分为$N=5$即6组, 即取$T_0=0$, $T_1=10$, $T_2=50$, $T_3=100$, $T_4=200$, $T_5=300$, $T_6=+\infty$, 误差精度取为0.001, 并且对$j\le 4$, 产品在$T_j$被删失的概率取为$j/5$, 而到$T_5$所有未失效的产品都被删失, 考虑在样本量$n=40$, 80, 120, 200, 500时重复模拟$s=1000$次, 若记第$k$次试验得到的估计为$\eta^{(k)}=(p_1^{(k)},\lambda_1^{(k)},\lambda_2^{(k)})$ $(k=1,\ldots,s)$, 则取最后的估计值及估计的标准差(Standard Deviation)分别为

$$\widehat{\mathrm{Mean}(\eta_j)}=\frac{1}{s}\sum_{k=1}^{s}\eta_j^{(k)}, \qquad \widehat{\mathrm{Std}(\eta_j)}=\sqrt{\frac{1}{s-1}\sum_{k=1}^{s}(\eta_j^{(k)}-\widehat{\mathrm{Mean}(\eta_j)})^2},$$

其中$\eta_j$表示$\eta$的第$j$个分量, 计算的相应结果见表1和表2.

表1 估计值(Mean)

| $\eta=(p_1,\alpha_1,\alpha_2,\lambda_1,\lambda_2)$ | $n$ | Mean | | | | |
|---|---|---|---|---|---|---|
| | | $\widehat{p}_1$ | $\widehat{\alpha}_1$ | $\widehat{\alpha}_2$ | $\widehat{\lambda}_1$ | $\widehat{\lambda}_2$ |
| $(0.4,0.8,0.5,0.06,0.004)$ | 40 | 0.3536 | 0.8743 | 0.4663 | 0.0723 | 0.0056 |
| | 80 | 0.4091 | 0.8025 | 0.4462 | 0.0691 | 0.0039 |
| | 120 | 0.4125 | 0.7995 | 0.4479 | 0.0688 | 0.0038 |
| | 200 | 0.4124 | 0.7979 | 0.4480 | 0.0686 | 0.0038 |
| | 500 | 0.4123 | 0.7966 | 0.4478 | 0.0680 | 0.0039 |

表2 估计的标准差(Std)

| $\eta=(p_1,\alpha_1,\alpha_2,\lambda_1,\lambda_2)$ | $n$ | Std | | | | |
|---|---|---|---|---|---|---|
| | | $\widehat{p}_1$ | $\widehat{\alpha}_1$ | $\widehat{\alpha}_2$ | $\widehat{\lambda}_1$ | $\widehat{\lambda}_2$ |
| $(0.4,0.8,0.5,0.06,0.004)$ | 40 | 0.0843 | 0.1758 | 0.0922 | 0.0230 | 0.0018 |
| | 80 | 0.0415 | 0.0696 | 0.0559 | 0.0079 | 6.0594E-4 |
| | 120 | 0.0285 | 0.0542 | 0.0294 | 0.0061 | 3.1630E-4 |
| | 200 | 0.0223 | 0.0416 | 0.0230 | 0.0047 | 2.3354E-4 |
| | 500 | 0.0140 | 0.0263 | 0.0145 | 0.0029 | 1.4653E-4 |

从表1和表2可以看到所提算法对于数据分组和右删失情形下混合广义指数分布有着很好的估计效率, 并且随着样本量的增大估计估计的标准误变得越来越小. 当然, 估计效果与初值点的选择有很大关系, 可选取多组不同的初值点比较估计效果. 另外, 从实际模拟来看, 估计效果受数据分组方式的影响不是很大.

## §5. 一组实际数据分析

下面分析一组医学数据来说明本文方法的实际应用. 心绞痛是冠状动脉供血不足, 心肌急剧的、暂时缺血与缺氧所引起的临床综合征, 常见于男性. 下表3中2418个患心绞痛的男性病人数据摘自[15]中Parker (1946)等人的工作, 生存时间是从诊断时间起按年计算的, 共有16个区间, 前15个区间长度均为一年, 即$I_j=(j-1,j]$, $j=1,2,\ldots,15$, $I_{16}=(15,\infty)$. 每个区间中死亡病例数和失去追踪病例数如表3所示.

该数据在[15]中是用乘积限的非参数方法估计生存函数和危险率函数的, 并得到结论: 诊断后的第1年死亡率最高, 从第1年末到第10年初死亡率基本上保持不变, 在$0.09-0.12$之间波动, 危险率函数在10年后一般比较高. 因此, 如果不考虑年龄、性别、种族因素, 活过一年的患者比刚诊断出的患者的预后情况好, 5年生存率是0.5193.

本文在此考虑2成分即$m=2$时的混合广义指数分布模型(1.1)分析这组数据, 用本文的算法得参数估计为$\widehat{p}_1=0.2504$, $\widehat{\alpha}_1=0.8532$, $\widehat{\alpha}_2=1.0451$, $\widehat{\lambda}_1=0.7499$, $\widehat{\lambda}_2=0.0869$, 则模

型(1.1)的生存函数和危险率函数分别是

$$\widehat{S}(x)=1-[0.2504\times(1-e^{-0.7499x})^{0.8532}+0.7496\times(1-e^{-0.0869x})^{1.0451}],\quad x\geq 0,\quad (5.1)$$

$$\widehat{h}(x)=\frac{0.1602\times(1-e^{-0.7499x})^{-0.1468}e^{-0.7499x}+0.0681\times(1-e^{-0.0869x})^{0.0451}e^{-0.0869x}}{1-[0.2504\times(1-e^{-0.7499x})^{0.8532}+0.7496\times(1-e^{-0.0869x})^{1.0451}]},$$
$$x\geq 0.\quad (5.2)$$

表3　男性心绞痛病人的生存数据

| Interval $I_j$ | Death numbers $D_j$ | Outfollowed numbers $W_j$ |
|---|---|---|
| 1 | 456 | 0 |
| 2 | 226 | 39 |
| 3 | 152 | 22 |
| 4 | 171 | 23 |
| 5 | 135 | 24 |
| 6 | 125 | 107 |
| 7 | 83 | 133 |
| 8 | 74 | 102 |
| 9 | 51 | 68 |
| 10 | 42 | 64 |
| 11 | 43 | 45 |
| 12 | 34 | 53 |
| 13 | 18 | 33 |
| 14 | 9 | 27 |
| 15 | 6 | 23 |
| 16 | 0 | 0 |

由(5.1)和(5.2)知拟合的生存函数图和危险率函数图如图1. 从图1看出, 危险率函数是单调递减函数, 前2年相对较大, 第1年为0.1739, 第2年为0.1305, 第5年为0.0912, 从第6年开始到第15年在0.0884 − 0.0863之间波动, 这说明心绞痛病的诊疗头几年比较关键, 早期危险率比较高. 此外, 根据拟合寿命模型(5.1), 得平均寿命是9.1770(年), 5年的生存概率是0.5026 (文献[15]中给出的结果是0.5193). 另一个常用的寿命指标是平均剩余寿命, 时刻$t$的平均剩余寿命公式为

$$\mu(t)=\frac{1}{S(t)}\int_t^{\infty}S(x)\mathrm{d}x,$$

计算得第1年的平均剩余寿命是10.3724(年), 第5年的平均剩余寿命是11.5283(年), 第10年的平均剩余寿命是11.5717(年). 因此, 如果不考虑年龄、性别、种族因素, 活过几年的患者

比刚诊断出的患者的平均剩余寿命要长些, 注意到本文所得平均剩余寿命相比较文献[15]有些偏高, 但这些数据分析所得结论都仅仅是提供一些参考, 实际中还要根据临床实践来作出更为合理的结论. 当然, 随着现代医学技术水平的不断提高和治疗心绞痛药物疗效的逐步改善, 患心绞痛病人的存活率和平均剩余寿命已大大提高.

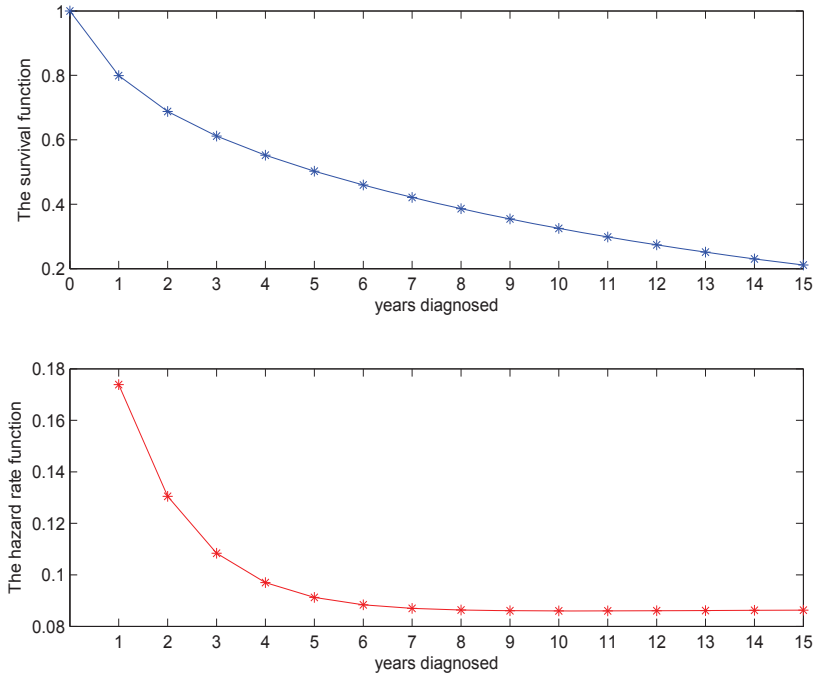


图1 男性心绞痛病人数据拟合的生存函数和危险率函数图

## 参 考 文 献

[1] Gupta, R.D. and Kundu, D., Generalized exponential distributions, *Australian & New Zealand Journal of Statistics*, **41**(1999), 173–188.

[2] Raqab, M.Z., Inferences for generalized exponential distribution based on record statistixs, *Journal of Statistical Planning and Inference*, **104**(2002), 339–350.

[3] Sarhan, A.M., Analysis of incomplete, censored data in competing risks models with generalized exponential distribution, *IEEE Transactions on Reliability*, **56**(2007), 132–138.

[4] Gupta, R.D. and Kundu, D., Generalized exponential distribution: existing results and some recent developments, *Journal of Statistical Planning and Inference*, **137**(2007), 3537–3547.

[5] Kundu, D. and Gupta, R.D., Generalized exponential distribution: Bayesian estimations, *Computational Statistics & Data Analysis*, **52**(2008), 1873–1883.

[6] Kundu, D. and Pradhan, B., Estimating the parameters of the generalized exponential distribution in presence of hybrid censoring, *Communications Statistics - Theory and Methods*, **38**(2009), 2030–2041.

[7] Chen, D.G. and Lio, Y.L., Parameter estimations for generalized exponential distribution under progressive type-I interval censoring, *Computational Statistics & Data Analysis*, **54**(2010), 1581–1591.

[8] Sultan, K.S., Ismail, M.A. and Al-Moisheer, A.S., Mixture of two inverse Weibull distribution: properties and estimation, *Computational Statistics & Data Analysis*, **51**(2007), 5377–5387.

[9] Jones, P.W. and Ashour, S.K., Bayesian estimation of the parameters of the mixed exponential distribution from censored samples, *Biometrical Journal*, **18**(1976), 633–637.

[10] Elsherpieny, E.A., Estimation of parameters of mixed generalized exponentionally distributions from censored type I samples, *Journal of Applied Sciences Research*, **3**(2007), 1696–1700.

[11] Pettitt, A.N., Re-weighted least squares estimation with censored and grouped data: an application of the EM algorithm, *Journal of the Royal Statistical Society: Series B*, **47**(1985), 253–260.

[12] Liu, L.P., Estimation of MLE for Weibull distribution with grouped and censored data, *Chinese Journal of Applied Probability and Statistics*, **17**(2001), 133–138.

[13] Liu, X., Chen, H. and Fei, H.L., Estimation of the parameters in the lognormal distribution with grouped and right-censored data, *Chinese Journal of Applied Probability and Statistics*, **24**(2008), 371–380.

[14] Dempster, A.P., Laird, N.M. and Rubin, D.B, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B*, **39**(1977), 1–38.

[15] Lee, E.T. and Wang, J.W., *Statistical Methods for Survival Data Analysis* (3rd Edition), Wiley-Interscience: Wiley, John and Sons, Incorporated, 2003.

# Parameter Estimation for a Mixture of Generalized Exponential Distributions under Grouped and Right-Censored Samples

Tian Yuzhu[1,2] Tian Maozai[2] Chen Ping[3]

([1]*School of Mathematics and Statistics, Tianshui Normal university, Tianshui, 741001*)

([2]*School of Statistics, Renmin University of China, Beijing, 100872*)

([3]*Department of Mathematics of Southeast University, Nanjing, 210096*)

Parameter estimation of mixed generalized exponential distribution model under grouped and right-censored data is considered by using EM algorithm in this paper. The estimation formulae are obtained and some simulations are presented to illustrate the proposed method. Finally, a set of medicine data is analyzed.

**Keywords:** Mixed generalized exponential distribution, grouped and right-censored data, EM algorithm.

**AMS Subject Classification:** 62N01.