应用概率统计 第二十四卷
第三期 2008年6月

Chinese Journal of Applied Probability
and Statistics Vol.24 No.3 Jun. 2008

二元极值分布混合模型的矩估计

尹 剑 史道济

(天津大学数学系; 南开大学天津大学刘徽应用数学中心, 天津, 300072)

摘 要

极值理论在各个领域得到了越来越多的关注和应用, 尤其是多元极值分布. 而矩估计是一种经典的参数估计方法, 计算简单且具有某些优良性, 本文给出边缘为标准指数分布的二元极值混合模型相关参数的矩估计及其渐近方差. 并将其与极大似然估计的渐近方差比较, 结果表明矩估计是一个较好的估计.

关键词: Copula, 二元极值分布, 混合模型, 矩估计, 渐近方差.

学科分类号: O213.

§1. 序 言

近些年, 多元极值模型在各种环境极值、风险度量等问题中得到了广泛应用. 参数估计是模型应用中首要解决的问题. [1]-[3]给出了多元logistic模型和嵌套logistic模型的矩估计和渐近方差阵. 但在实际应用中, [4]发现用二元极值分布混合模型可以得到更好的结果.

本文考虑的是二元极值混合模型相关参数的矩估计及渐近方差. §2介绍了copula与联合分布的关系, §3给出了二元极值分布混合模型的各阶矩, §4得到了模型中相关参数矩估计的渐近方差, §5比较相关参数矩估计的渐近方差与极大似然估计的渐近方差.

§2. Copula

联合分布函数包含两方面的信息, 一是变量的边缘分布信息, 另一是变量间相关结构的信息. 设随机向量(X,Y)的联合分布函数为$F(x,y)$, 那么$F_1(x)=F(x,+\infty)$, $F_2(y)=F(+\infty,y)$分别是X,Y的边缘分布, 即由联合分布函数容易得到变量的边缘分布函数, 因此在联合分布中除去边缘分布的信息后, 就剩下相关结构的信息了. 如果存在函数C, 使

$$F(x,y)=C(F_1(x),F_2(y)),$$

则称C是分布函数F的copula, 有时也称C为随机向量(X,Y)的copula, 且记为$C_{X,Y}$.

本文2005年1月10日收到, 2005年6月14日收到修改稿.

反过来, 如果$F_1(x), F_2(y)$为连续分布函数, 令$U = F_1(X)$, $V = F_2(Y)$, 它们服从$[0,1]$上均匀分布, 则(U,V)的联合分布函数为

$$\mathsf{P}(U \le u, V \le v) = F(F_1^{-1}(u), F_2^{-1}(v)) = C(u,v),$$

其中F_j^{-1}是F_j的反函数, 有时也称它为分布的分位数函数, 因此C可以看作是边缘分布为区间$[0,1]$上均匀分布U,V的联合分布函数. 在这个意义上, C反映了X,Y之间的关联, 且这种关联与边缘分布无关. Copula描述了联合分布函数与边缘分布函数之间的关系, [5]最先给出这个概念.

一般地, 对任意$t>0$, 如果copula C满足

$$C(u^t, v^t) = C^t(u,v),$$

则称C为二元极值copula. 可以将C表示为依赖于一个函数的形式

$$C(u,v) = \exp\Big\{\log(uv) A\Big(\frac{\log v}{\log(uv)}\Big)\Big\}. \tag{2.1}$$

称C是由A生成的极值copula, 或

$$C(u,v) = \exp\{-V(-1/\log u, -1/\log v)\}, \tag{2.2}$$

其中, 函数A是定义在$[0,1]$上的Pickands相关函数, 满足

$$\max(t, 1-t) \le A(t) \le 1, \qquad 0 \le t \le 1.$$

而V称为指数测度(Exponent measure), 是-1阶齐次函数[6], 满足

$$V(ax, ay) = a^{-1} V(x,y).$$

(2.1)只适用于二元极值, 而(2.2)对一般的多元极值也成立.

§3. 二元极值混合模型的各阶矩

二元极值混合模型的相关函数为:

$$A(t) = \theta t^2 - \theta t + 1, \qquad 0 \le t \le 1,$$

其中$0 \le \theta \le 1$为模型的相关参数. 于是由(2.1)可以很容易得到二元极值分布混合模型的分布函数. 对于一般的边缘分布函数不易得到模型相关系数的显式表达式, 但通过某些变换, 可将一般的边缘分布函数变换为标准指数分布, 因此为简单起见且不失一般性, 这里考虑

边缘为标准指数分布的二元极值混合模型. 由混合模型的相关函数和copula, 易得模型的分布函数为:

$$G(x,y)=\exp\Big\{-(x+y)+\theta\frac{xy}{x+y}\Big\},\qquad x>0,\ y>0,\tag{3.1}$$

其中相关参数$0\le\theta\le 1$.

通过变换[8]

$$\begin{cases} s=x+y-\theta\cdot[xy/(x+y)];\\ t=x/(x+y),\end{cases}\tag{3.2}$$

或写成:

$$\begin{cases} x=st/A(t);\\ y=s(1-t)/A(t).\end{cases}\tag{3.3}$$

可以得到(3.1)的联合密度函数的简单形式为:

$$g(x,y)=\frac{\partial^2 G(x,y)}{\partial x\partial y}=\frac{A^2(t)}{s}e^{-s}W(s,t),\tag{3.4}$$

其中

$$\begin{aligned} W(s,t)&=p_1(t)s+p_2(t),\\ p_1(t)&=\frac{4-\theta}{A^2(t)}-\frac{4}{A(t)}+1,\\ p_2(t)&=\frac{2}{A(t)}-2,\end{aligned}$$

变换(3.2)的Jacobian:

$$\frac{\partial(s,t)}{\partial(x,y)}=\frac{A^2(t)}{s},$$

于是, 得到S,T的联合分布密度为

$$f(s,t)=e^{-s}[p_1(t)s+p_2(t)],\qquad s>0,\ 0<t<1.$$

由此容易得到S的边缘分布密度为

$$q(s)=[(1-\beta)s+\beta]e^{-s},\qquad s>0,$$

其中

$$\beta=\int_0^1 p_2(t)\mathrm{d}t=\frac{8}{\sqrt{\theta(4-\theta)}}\mathrm{arctg}\sqrt{\frac{\theta}{4-\theta}}-2,\tag{3.5}$$

即S的分布是$\Gamma(s,1)$与$\Gamma(s,2)$的混合分布, 这里$\Gamma(s,k)$表示参数为k的Gamma分布密度:

$$\Gamma(s,k)=s^{k-1}e^{-s}/\Gamma(k),\qquad s>0.$$

而T的边缘分布密度为

$$p(t)=p_1(t)+p_2(t)=\frac{4-\theta}{A^2(t)}-\frac{2}{A(t)}-1,\qquad 0<t<1.$$

由于模型(3.1)中变量是对称可交换的, 为了简便, 记μ_{ab}为$(a+b)$阶矩, 即

$$\mu_{ab}=\mathsf{E}(X-\mathsf{E}X)^a(Y-\mathsf{E}Y)^b.$$

由标准指数分布的性质可知$\mathsf{E}(X)=\mathsf{E}(Y)=1$, 注意到变换(3.3), 我们有

$$\mu_{ab}=\mathsf{E}\Big[\frac{st}{A(t)}-1\Big]^a\Big[\frac{s(1-t)}{A(t)}-1\Big]^b,$$

因此, 若要计算μ_{ab}, 则只需计算形如$s^it^k/A^j(t)$, (i,j,k为整数)的数学期望即可. 经复杂的积分计算可以得到:

$$\begin{aligned}
&\mathsf{E}\Big[\frac{st}{A(t)}\Big]^k=k!,\\
&\mathsf{E}\Big[\frac{s^it}{A^j(t)}\Big]=\frac{1}{2}\mathsf{E}\Big[\frac{s^i}{A^j(t)}\Big],\\
&\mathsf{E}\Big[\frac{s^it^2}{A^j(t)}\Big]=\frac{1}{\theta}\mathsf{E}\Big[\frac{s^i}{A^{j-1}(t)}\Big]+\Big(\frac{1}{2}-\frac{1}{\theta}\Big)\mathsf{E}\Big[\frac{s^i}{A^j(t)}\Big],\\
&\mathsf{E}\Big[\frac{s^it^3}{A^j(t)}\Big]=\frac{3}{2\theta}\mathsf{E}\Big[\frac{s^i}{A^{j-1}(t)}\Big]+\Big(\frac{1}{2}-\frac{3}{2\theta}\Big)\mathsf{E}\Big[\frac{s^i}{A^j(t)}\Big],
\end{aligned}$$

所以各阶矩的计算就都转化为$\mathsf{E}[s^i/A^j(t)]$的计算, 根据(S,T)的联合密度可得

$$\mathsf{E}\Big[\frac{s^i}{A^j(t)}\Big]=\int_0^1\frac{1}{A^j(t)}[(i+1)!p_1(t)+i!p_2(t)]\mathrm{d}t,$$

这个积分是很容易计算的. 经过化简整理得到:

$$\mu_{20}=1,\qquad \mu_{30}=2,\qquad \mu_{40}=9,\qquad \mu_{11}=\frac{\beta+\theta}{4-\theta},\tag{3.6}$$

$$\begin{aligned}
&\mu_{12}=-1-\frac{14+4\beta}{4-\theta}+\frac{6(4+\beta)}{(4-\theta)^2},\\
&\mu_{13}=3\Big[\frac{3\beta}{16\theta}-1+\frac{19\beta+64}{16(4-\theta)}-\frac{21\beta+64}{4(4-\theta)^2}+15\frac{4+\beta}{(4-\theta)^3}\Big],
\end{aligned}$$

$$\mu_{22}=1-\frac{3\beta}{4\theta}+\frac{13\beta+32}{4(4-\theta)}-\frac{27\beta+88}{(4-\theta)^2}+60\frac{4+\beta}{(4-\theta)^3}.\tag{3.7}$$

§4. 相关参数矩估计的渐近方差

由(3.6)可以得出边缘为指数分布的二元极值混合模型的相关系数

$$\rho(\theta)=\frac{\beta+\theta}{4-\theta},\tag{4.1}$$

其中β如(3.5)所示. 通过(4.1)不能得到相关参数θ的矩估计的显式表达式, 我们用记号$\widehat{\theta}=\phi(r)$来表示由(4.1)反解得到的参数θ的矩估计, 其中r为样本相关系数, 这样做并不影响对其渐近方差的计算. Stuart, A. & Ord, J.K.给出了一种求方差的近似方法[9]: 考虑k维随机向量$\xi=(\xi_1,\cdots,\xi_k)$的函数$f(\xi)=f(\xi_1,\cdots,\xi_k)$. 假设$\xi_1,\cdots,\xi_k$的方差和协方差都是$n^{-1}$阶的, 如果$\mathsf{E}(\xi_i)=a_i$, $i=1,\cdots,k$, $f_i'(a)$表示$\partial f(\xi)/\partial\xi_i$在$a=(a_1,\cdots,a_k)$处的值, 则由函数$f(\xi)$的Taylor展开式

$$f(\xi)=f(a)+\sum_{i=1}^{k}f_i'(a)(\xi_i-a_i)+O(n^{-1})$$

可得

$$\begin{aligned}\mathsf{E}[f(\xi)] &= f(a)+O(n^{-1}),\\ \mathsf{Var}\,[f(\xi)] &= \mathsf{E}\Big\{\sum_{i=1}^{k}f_i'(a)(\xi_i-a_i)\Big\}^2+O(n^{-1})\\ &= \sum_{i=1}^{k}[f_i'(a)]^2\mathsf{Var}\,(\xi_i)+\sum_{i\neq j}f_i'(a)f_j'(a)\mathsf{Cov}\,(\xi_i,\xi_j)+O(n^{-3/2}),\end{aligned}\tag{4.2}$$

以及两个函数$g(\xi)$, $h(\xi)$间的协方差

$$\mathsf{Cov}\,[g(\xi),h(\xi)]=\sum_{i=1}^{k}g_i'(a)h_i'(a)\mathsf{Var}\,(\xi_i)+\sum_{i\neq j}g_i'(a)h_j'(a)\mathsf{Cov}\,(\xi_i,\xi_j)+O(n^{-3/2}).\tag{4.3}$$

因此若要得到$\widehat{\theta}$的渐近方差, 只要求出样本相关系数的方差即可. 记样本相关系数

$$r=\frac{m_{11}}{\sqrt{m_{20}m_{02}}},$$

其中

$$m_{11}=\sum_{i=1}^{n}(X_i-\overline{X})^2(Y_i-\overline{Y})^2,\qquad m_{20}=\sum_{i=1}^{n}(X_i-\overline{X})^2,\qquad m_{02}=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$$

为样本2阶矩,

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\qquad \overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

是样本均值. 由

$$\mathsf{Var}\,\Big(\frac{x_1}{x_2}\Big)=\Big\{\frac{\mathsf{E}(x_1)}{\mathsf{E}(x_2)}\Big\}^2\Big\{\frac{\mathsf{Var}\,x_1}{\mathsf{E}^2(x_1)}+\frac{\mathsf{Var}\,x_2}{\mathsf{E}^2(x_2)}-\frac{2\mathsf{Cov}\,(x_1,x_2)}{\mathsf{E}(x_1)\mathsf{E}(x_2)}\Big\}$$

和(4.2), (4.3)可知

$$\begin{aligned}\mathsf{Var}\,r &= \rho^2\Big\{\frac{\mathsf{Var}\,m_{11}}{\mu_{11}^2}+\frac{1}{4}\Big(\frac{\mathsf{Var}\,m_{20}}{\mu_{20}^2}+\frac{\mathsf{Var}\,m_{02}}{\mu_{02}^2}+\frac{2\mathsf{Cov}\,(m_{20},m_{02})}{\mu_{20}\mu_{02}}\Big)\\ &\quad -\frac{\mathsf{Cov}\,(m_{11},m_{02})}{\mu_{11}\mu_{02}}-\frac{\mathsf{Cov}\,(m_{11},m_{20})}{\mu_{11}\mu_{20}}\Big\},\end{aligned}$$

再由样本各阶矩的方差和协方差计算公式:

$$\begin{aligned}\mathsf{Var}\, m_{r,s} &= \frac{1}{n}(\mu_{2r,2s}-\mu_{r,s}^2+r^2\mu_{20}\mu_{r-1,s}^2+s^2\mu_{02}\mu_{r,s-1}^2\\&\quad+2rs\mu_{11}\mu_{r-1,s}\mu_{r,s-1}-2r\mu_{r+1,s}\mu_{r-1,s}-2s\mu_{r,s+1}\mu_{r,s-1}),\\ \mathsf{Cov}\,(m_{r,s},m_{u,v}) &= \frac{1}{n}(\mu_{r+u,s+v}-\mu_{r,s}\mu_{u,v}+r\mu\mu_{20}\mu_{r-1,s}\mu_{u-1,v}+sv\mu_{02}\mu_{r,s-1}\mu_{u,v-1}\\&\quad+rv\mu_{11}\mu_{r-1,s}\mu_{u,v-1}+su\mu_{11}\mu_{r,s-1}\mu_{u-1,v}-u\mu_{r+1,s}\mu_{u-1,v}\\&\quad-v\mu_{r,s+1}\mu_{u,v-1}-r\mu_{r-1,s}\mu_{u+1,v}-s\mu_{r,s-1}\mu_{u,v+1})\end{aligned}$$

以及二元混合模型变量的可交换性, 得到

$$\begin{aligned}\mathsf{Var}\, r &= \frac{\rho^2}{n}\Big\{\frac{\mu_{22}}{\mu_{11}^2}+\frac{\mu_4+\mu_{22}}{2\mu_2^2}-\frac{2\mu_{31}}{\mu_{11}\mu_2}\Big\}\\ &= \frac{1}{n}\Big\{\frac{9}{2}\rho^2+\Big(1+\frac{\rho^2}{2}\Big)\mu_{22}-2\rho\mu_{31}\Big\},\end{aligned}\tag{4.4}$$

代入(3.7)中的结果, 则有

$$\begin{aligned}n\mathsf{Var}\, r &= -\frac{3\beta}{128\theta}(4+\beta)(8+\beta)-\frac{1}{128(4-\theta)}(3\beta^3+36\beta^2-928\beta+2560)\\&\quad-\frac{1}{32(4-\theta)^2}(3\beta^3+196\beta^2+3424\beta+8832)\\&\quad+\frac{4+\beta}{8(4-\theta)^3}(13\beta^2+552\beta+3040)\\&\quad-\frac{1}{2(4-\theta)^4}(27\beta^3+604\beta^2+3536\beta+3104)\\&\quad+\frac{30}{(4-\theta)^5}(\beta^3+12\beta^2+460\beta+640),\\ n\mathsf{Var}\,\widehat{\theta} &= \frac{\theta^2(4-\theta)^4}{4(\beta\theta+3\theta-\beta)^2}\mathsf{Var}\, r.\end{aligned}\tag{4.5}$$

给定不同的θ, 由(4.5)很容易得到矩估计的渐近方差$n\mathsf{Var}\,\widehat{\theta}$. 表1给出了$n=100$时渐近方差的数值计算结果.

§5. 与极大似然估计渐近方差的比较

本节我们比较二元极值分布混合模型(3.1)的相关参数矩估计与极大似然估计的渐近方差. 我们知道, 极大似然估计的渐近方差阵就是Fisher信息阵的逆矩阵. 由(3.4)可以得到单个观测值的对数似然函数:

$$l=\log g(x,y)=2\log A(t)-\log s-s+\log[p_1(t)s+p_2(t)].$$

而得分统计量函数为:

$$\frac{\partial l}{\partial\theta}=\frac{1}{\theta}\Big[\Big(\frac{1}{A(t)}-\frac{1}{W(s,t)}-1\Big)s+1-\frac{p_2(t)s}{W(s,t)}\Big].$$

文献[10]给出了二元极值分布混合模型的Fisher信息阵, 其中

$$\begin{aligned}\theta^2\mathsf{E}\Big(\frac{\partial l}{\partial\theta}\Big)^2 &= 4-2\frac{\beta-2}{4-\theta}-\frac{4(1+\sqrt{\theta})^3}{\theta(2+\sqrt{\theta})^2}\log(1+\sqrt{\theta})\\ &\quad -\frac{4(1-\sqrt{\theta})^3}{\theta(2-\sqrt{\theta})^2}\log(1-\sqrt{\theta})+\mathsf{E}\Big[\frac{p_2^2(t)(p_2(t)+1)^2}{p_1^2(t)W^2(s,t)}\Big],\end{aligned}\tag{5.1}$$

其中$\mathsf{E}(\cdot)$表示在密度(3.4)下的期望, 而$\mathsf{E}\{[p_2^2(t)(p_2(t)+1)^2]/[p_1^2(t)W^2(s,t)]\}$可以由数值计算方法很容易得到结果. 则$\mathsf{E}(\partial l/\partial\theta)^2$的倒数就是参数$\theta$的极大似然估计的渐近方差.

给定不同的θ, 由(4.5)很容易得到矩估计的渐近方差$n\mathsf{Var}\,\widehat{\theta}$. 表1给出了$n=100$时矩估计的渐近方差数值计算结果, 括号中的数值是$n=100$时, 由(5.1)计算得到的极大似然估计的渐近方差.

表1 不同θ值对应$\widehat{\theta}$的渐近方差值

θ	0.05	0.1	0.2	0.3	0.4
$\mathsf{Var}\,\widehat{\theta}$	0.0870 (0.0731)	0.0839 (0.0391)	0.0777 (0.0222)	0.0714 (0.0167)	0.0649 (0.0142)
θ	0.5	0.6	0.7	0.8	0.9
$\mathsf{Var}\,\widehat{\theta}$	0.0582 (0.0130)	0.0514 (0.0125)	0.0444 (0.0126)	0.0374 (0.0135)	0.0302 (0.0160)

虽然我们没有得到相关参数矩估计的显式表示, 但通过数值计算可以很容易从(4.1)得到估计值. 由大样本下矩估计的渐近正态性可知, 该估计量是渐近无偏的, 且从表1可以看出对于所有的$0<\theta<1$, 矩估计的渐近方差与极大似然估计的渐近方差相差不大. 由此对于该模型来说, 矩估计是一个较好的估计.

参 考 文 献

[1] Shi, D.J., Moment estimation for multivariate extreme value distributon, *J. Appl. Math.*, **B10**(1995), 61–68.

[2] Shi, D.J. and Zhou, S.S., Moment estimation for multivariate extreme value distribution, *Appl. Inst. Statist. Math.*, **51**(1999), 253–264.

[3] 史道济, 孙炳堃, 嵌套Logistic模型的矩估计, 系统工程理论与实践, **21(1)**(2001), 53–60.

[4] Sheng, Y., The gumbel mixed model applied to storm frequency analysis, *Water Resources Management*, **14(5)**(2000), 377–389.

[5] Sibuya, M., Bivariate extreme statistics, *Ann. Inst. Statist. Math.*, **11**(1960), 195–210.

[6] Pickands, J., Multivariate extreme value distributions, *Proc. 43rd Session of the ISI*, Buenos Aires, **49**(1981), 859–878.

[7] Tiago de Oliveira, J., Bivariate models for extreme: statistical decision, In *Statistical Extreme and Applications*, Dordrecht: Reidel, 1984, 131–153.

[8] 史道济, 二元极值分布的一个性质, 应用概率统计, **19(1)**(2003), 49-54.

[9] Stuart, A. and Ord, J.K., *Kendall's Advanced Theory of Statistics*, Vol.1: Distribution Theory (5-th edn.), Charles Griffin, 1987.

[10] 史道济, 尹剑, 二元极值分布混合模型的Fisher信息阵, 已投工程数学学报.

Moment Estimation for a Bivariate Extreme Value Distribution in Mixed Model

Yin Jian Shi Daoji

(*Department of Mathematics, Tianjin University;*)

(*Nankai University and Tianjin University LiuHui Applied Mathematics Center, Tianjin, 300072*)

Extreme value theory has been wildly applied in many fields, especially the multivariate extreme value distributions. Moment estimation is a classical estimation method because of its simple calculations. The paper considers the bivariate extreme value distribution in mixed model with exponential margins. The estimator and asymptotic variance of the dependence parameter are given. We also compare moment estimation with a maximum likelihood estimation in finite sample size. The results indicate that moment estimation is good for all practical purposes.

Keywords: Copula, bivariate extreme value distribution, mixed model, moment estimation, asymptotic variance.

AMS Subject Classification: 62G32.