第 35 卷 第 3 期 应 用 数 学 学 报 Vol. 35 No. 3
2012 年 5 月 ACTA MATHEMATICAE APPLICATAE SINICA May, 2012

二项 - 广义 Pareto 复合极值分布模型的统计推断 *

张香云

(浙江农林大学天目学院, 临安 311300)

(E-mail: zlzhangxiangyun@163.com)

程维虎

(北京工业大学应用数理学院, 北京 100124)

(E-mail: chengweihu@bjut.edu.cn)

摘 要 极值理论主要研究小概率、大影响的极端事件. 当前, 复合极值分布已经广泛应用于水文、气象、地震、保险、金融等领域. 本文以极值类型定理和 PBDH 定理为理论依据, 构建了二项 - 广义 Pareto 复合极值分布模型; 使用概率加权矩方法, 对所建立的复合模型推导参数估计式; 利用计算机模拟, 得到了 Kolmogorov-Smirnov (简称 KS) 检验统计量的临界值.

关键词 广义 Pareto 分布; 二项分布; 概率加权矩; Anderson-Darling(AD) 检验; KS 检验; 次序统计量; 随机模拟

MR(2000) 主题分类 62F07; 62F12

中图分类 O213.2

1 引言

极值事件虽然很少发生, 但其一旦发生, 将会产生空前重大的影响. 自然环境中的极值事件, 如百年一遇的洪水、地震、干旱或飓风等, 常常打破自然界相对平衡的和谐状态, 对人类的生产和生活造成巨大损失. 极值理论就是专门研究这些很少发生、而一旦发生却具有重大影响的随机事件. 该理论的核心问题是对极值事件的统计分析, 提供一个优良稳健的渐近模型, 用于对分布的尾部进行建模, 并评估极值事件的风险.

德国统计学家 L. von Bortkiewicz 于 1922 年首次提出极值问题 [1], 并研究正态分布

本文 2010 年 10 月 31 日收到. 2011 年 10 月 19 日收到修改稿.

* 2010 年北京市教委科技面上 (KM201010005006), 北京工业大学人才强教深化计划 –211 工程 – 服务北京优秀团队 (00600054R0001) 资助项目.

样本极差，给出了正态总体样本的最大值分布； 1923 年， R. von Mises[2] 研究了样本最大值的期望；同年， Dodd E L[3] 讨论了一般分布的样本最大值问题； 1927 年， Fréchet M[4] 提出了最大值稳定原理，提出了“来自不同分布，但有某种共同性质的最大值可以有相同的渐近分布”的理论； 1928 年， Fisher R A 与 Tippet L H C[5] 明确提出了极值分布的三种类型，即极值类型定理； 1958 年， Gumbel 的著作 [6] 奠定了极值理论的基础，并成为随机变量极端变异性的建模工具.

最初，极值理论只是研究独立同分布随机变量的最大值或最小值的渐近性质. 但从大量的数据中仅选用最大值或最小值，会舍弃许多有价值的信息. 之后，极值理论主要研究选取某界限以上的数据进行分析的方法，即所谓的 POT(peaks over threshold) 法. 20 世纪 70 年代， Balkeman, Haan[7], Pickands[8] 相继证明了分布函数之所以属于广义极值分布的吸引场，是因为分布的某界限值以上的数据可由广义 Pareto 分布 (简称 GPD) 近似取得. 从此，基于极值理论的 GPD 成为极值理论研究的主要方向之一，并得到广大学者的认可，进而得以广泛应用.

1968 年， Feller[9] 首先提出了复合极值的概念. 随着极值理论的深入研究，复合分布已经广泛应用于水文、气象、地震、交通、环境与工程保险、金融等领域. 本文首先以极值类型定理和 PBDH 定理为依据，构建了二项 - 广义 Pareto 复合极值分布模型，并对所建立的模型使用概率加权矩方法推导模型参数的估计式；其次，利用计算机模拟计算提出模型的 AD 检验统计量和 KS 检验统计量的临界值.

2　广义 Pareto 分布模型与极值理论

GPD 首先由 Pickands[8] 在 1975 年引入到水文、气象学的研究. 后来， Hosking[10] 等进一步发展了该模型的应用. 当前，该分布已被广泛应用于极值分析、拟合保险损失、可靠性研究及金融风险管理等领域.

2.1　广义 Pareto 模型

若随机变量 X 有如下分布函数

$$G(x,\mu,\sigma,\delta)=\begin{cases}1-\left(1+\delta\dfrac{x-\mu}{\sigma}\right)^{-\frac{1}{\delta}}, & \delta\neq 0,\\ 1-\exp\left(\dfrac{x-\mu}{\sigma}\right), & \delta=0,\end{cases}\qquad x>\mu,\ \ 1+\delta\frac{x-\mu}{\sigma}>0, \tag{2.1}$$

其中 $-\infty<\mu<\infty$ 是位置参数， $\sigma>0$ 是刻度参数， $-\infty<\delta<\infty$ 是形状参数，则称随机变量 X 服从三参数广义 Pareto 分布.

当 $\delta\geq 0$ 时， $x\geq\mu$; 当 $\delta<0$ 时， $\mu\leq x\leq\mu-\sigma/\delta$. 由 (2.1) 式，得 $G(x;\mu,\sigma,\delta)$ 的反函数

$$x=\mu+\frac{\sigma}{\delta}[(1-G)^{-\delta}-1]. \tag{2.2}$$

2.2 极值类型定理

设 $X_1, X_2, \cdots, X_n$ 是抽自总体分布函数为 $F(x)$ 的随机样本，令

$$M_n = \max\{X_1, X_2, \cdots, X_n\},$$

则有如下定理.

定理 2.1 (Fisher-Tippett 极值类型定理)[5] 若存在常数列 $a_n > 0, b_n$, 使得

$$\lim_{n\to\infty} \Pr\Big(\frac{M_n - b_n}{a_n} \le x\Big) = H(x), \qquad -\infty < x < \infty$$

成立，则非退化分布函数 $H(x)$ 必属下列三种类型之一：

$$H_1(x) = \exp\{-\mathrm{e}^{-x}\} \text{ 为 Gumbel 分布;}$$

$$H_2(x) = \begin{cases} \exp\{-x^{-\alpha}\}, & x > 0, \\ 0, & x \le 0, \end{cases} \quad \alpha > 0 \text{ 为 Fr echet 分布;}$$

$$H_3(x) = \begin{cases} \exp\{-(-x)^{\alpha}\}, & x \le 0, \\ 1, & x > 0, \end{cases} \quad \alpha > 0 \text{ 为 Weibull 分布.}$$

且经变换，三种类型的分布有统一形式

$$H(x;\mu,\sigma,\delta) = \exp\Big\{-\Big(1+\delta\frac{x-\mu}{\sigma}\Big)^{-\frac{1}{\delta}}\Big\}, \qquad 1+\delta\frac{x-\mu}{\sigma} > 0,$$

这里 $H(\cdot)$ 被称为广义极值分布.

定义 2.1[11] 对给定的分布函数 $F(x)$, 若存在常数列 $a_n > 0,\ b_n$, 使得

$$\lim_{n\to\infty} F^n(a_n x + b_n) = F(x).$$

则称分布函数 $F(x)$ 是最大值稳定的 (max-stable).

可见：分布函数 $F(x)$ 是最大值稳定的，其充要条件是 $F(x)$ 是上述 $H(x)$ 的三种形式之一.

2.3 超阈值模型

假定 $X_1, X_2, \cdots, X_n$ 是抽自总体分布函数 $F(x)$ 未知的随机样本，设 u 为某一充分大的阈值，超过 u 的样本个数为 k, 为方便起见，不妨用 $X_1, X_2, \cdots, X_k$ 表示超过阈值的观测值，超过部分记成 $Y_i = X_i - u,\ i = 1, \cdots, k$, 分布函数为 $F_u(y)$. 令 x^* 表示分布 $F(x)$ 的密度函数支撑集的右端点，即 $x^* = \sup\{x : -\infty < x < \infty,\ F(x) < 1\},\ x^*$ 可能为无穷. $F_u(y)$ 的表达式为

$$F_u(y) = P(X - u \le y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)} = \frac{F(x) - F(u)}{1 - F(u)}. \tag{2.3}$$

定义 2.2[11]　设随机变量 X 的分布函数为 $F(x), x^*$ 表示分布 $F(x)$ 的密度函数支撑集的右端点， X 超过阈值 u 的超出量分布 (简称超量分布) 为 $F_u(x)$, 如果存在广义 Pareto 分布 $G(x;\mu,\sigma,\delta)$, 使得

$$\lim_{u\to x^*} F_u(x) = G(x;\mu,\sigma,\delta),$$

则称分布函数 $F(x)$ 属于广义 Pareto 分布 $G(x;\mu,\sigma,\delta)$ 的 POT 吸引场.

定理 2.2　(PBDH 定理, Balkema 和 De Haan, Pickands)[7,8]　若存在常数列 $a_n>0$, b_n 使 $u\longrightarrow x^*$ 时， $F_u(a_nx+b_n)$ 有连续的极值分布，则

$$\lim_{u\to x^*} F_u(x) = G(x;\mu,\sigma_u,\delta)$$

对于某个 δ 和 σ_u 成立.

由此可见： GPD 可作为高阈值超量的近似分布.

3　复合极值模型

随着极值分布在水文、海洋、气象、地震以及可靠性等领域内越来越广泛的应用，人们开始注意到应如何正确、合理地使用这种理论，才能得到更满意的结果.

3.1　二项 - 广义 Pareto 复合模型

在考虑如洪水、波高、地震的年最大值分布中，由于每年可利用的原始资料并非一个常数，而是一个随机变量. [9] 提出了复合极值分布的理论，即：假设 $X_1,X_2,\cdots,X_N$ 是来自分布函数为 $G(x)$ 的总体 X 的一个样本，这里的 N 是与 X 独立的随机变量，且

$$\Pr(N=k)=p_k,\qquad \sum_k p_k=1.$$

令 $\zeta=\max\{X_1,X_2,\cdots,X_N\}$, 则 ζ 的分布函数为 $F_\zeta(x)=\Pr(\zeta\le x)=\sum\limits_k p_k[G(x)]^k$, 称 $F_\zeta(x)$ 为由 N 和 X 构成的复合极值分布.

然而实际问题中的 N 通常取值有限，服从二项分布，可以预测，预测方法参见 [12] 的附录二. 本文讨论 N 服从参数为 (m,p) 的二项分布，其中 m 为正整数， p 为常数，$0<p<1$. X 服从广义 Pareto 分布，由此组成的复合极值分布函数仍用 $F(x)$ 表示，则二项 - 广义 Pareto 复合分布函数为

$$F(x)=\sum_{k=0}^{m} C_m^k p^k(1-p)^{m-k}G^k(x)=[pG(x)+1-p]^m, \tag{3.1}$$

其中 $G(\cdot)$ 由 (2.1) 式给出，代入到 (3.1) 式，得

$$F(x)=\Big[1-p\Big(1+\delta\frac{x-\mu}{\sigma}\Big)^{-\frac{1}{\delta}}\Big]^m. \tag{3.2}$$

从而

$$x = \mu - \sigma/\delta + (\sigma/\delta)p^{\delta}(1 - F^{1/m})^{-\delta} \tag{3.3}$$

是 $F(X)$ 的反函数.

3.2 二项 - 广义 Pareto 复合分布的参数估计

常用的参数估计方法有极大似然法、复合矩法和概率加权矩法 (Probability Weighted Moment, 简记 PWM). 本文用 PWM 给出二项 - 广义 Pareto 复合极值分布的参数估计.

设随机变量 X 的分布函数为 $F(x)$, 其概率加权矩定义为

$$M_{q,r,s} = E\{[X(F)]^q[F(X)]^r[1 - F(X)]^s\} = \int_0^1 x^q[F(x)]^r[1 - F(x)]^s \,\mathrm{d}F(x),$$

其中 $X(F)$ 是 $F(X)$ 的反函数.

取 $q = 1,\ r = 0,\ s = 0$, 当 $0 < \delta < 1$ 时, 得

$$\begin{aligned} M_{1,0,0} = E(X) =& \mu - \sigma/\delta + (\sigma/\delta)p^{\delta}E(1 - F^{1/m})^{-\delta} \\ =& \mu - \sigma/\delta + (\sigma/\delta)p^{\delta}mB(m, 1 - \delta); \end{aligned} \tag{3.4}$$

这里 $B(\cdot,\cdot)$ 为 Beta 函数.

取 $q = 1,\ r = 1,\ s = 0$, 当 $0 < \delta < 1$ 时, 得

$$\begin{aligned} M_{1,1,0} =& E(XF) = (\mu - \sigma/\delta)E(F) + (\sigma/\delta)p^{\delta}E[F(1 - F^{1/m})^{-\delta}] \\ =& 0.5(\mu - \sigma/\delta) + (\sigma/\delta)p^{\delta}mB(2m, 1 - \delta); \end{aligned} \tag{3.5}$$

取 $q = 1,\ r = 0.5,\ s = 0$, 当 $0 < \delta < 1$ 时, 得

$$\begin{aligned} M_{1,0.5,0} =& E(XF^{0.5}) = (\mu - \sigma/\delta)E(XF^{0.5}) + (\sigma/\delta)p^{\delta}E[F^{0.5}(1 - F^{1/m})^{-\delta}] \\ =& (2/3)(\mu - \sigma/\delta) + (\sigma/\delta)p^{\delta}mB(1.5m, 1 - \delta). \end{aligned} \tag{3.6}$$

由 (3.4) 至 (3.6) 式, 得

$$\frac{2M_{1,1,0} - M_{1,0,0}}{1.5M_{1,0.5,0} - M_{1,0,0}} = \frac{2B(2m, 1 - \delta) - B(m, 1 - \delta)}{1.5B(1.5m, 1 - \delta) - B(m, 1 - \delta)}.$$

易见: 当 m 已知时, 上式右端只是 δ 的函数. 分别用 $M_{1,0,0}, M_{1,1,0}, M_{1,0.5,0}$ 的加权矩估计

$$\widetilde{\beta_1} = \overline{x}, \qquad \widetilde{\beta_2} = [n(n+1)]^{-1}\sum_{i=1}^{n} ix_{(i)}, \qquad \widetilde{\beta_3} = (n\sqrt{n+1})^{-1}\sum_{i=1}^{n}\sqrt{i}x_{(i)}$$

代替 $M_{1,0,0}, M_{1,1,0}, M_{1,0.5,0}$, 得如下估计方程

$$\frac{2\widetilde{\beta_2} - \widetilde{\beta_1}}{1.5\widetilde{\beta_3} - \widetilde{\beta_1}} = \frac{2B(2m, 1 - \delta) - B(m, 1 - \delta)}{1.5B(1.5m, 1 - \delta) - B(m, 1 - \delta)}. \tag{3.7}$$

由 (3.7) 式得到 δ 的估计 $\widehat{\delta}$, 再将 $\widehat{\delta}$ 代入到下列方程

$$\begin{cases} \widetilde{\beta_1} = \mu - \sigma/\delta + (\sigma/\delta)p^{\delta}mB(m, 1-\delta), \\ \widetilde{\beta_2} = 0.5(\mu - \sigma/\delta) + (\sigma/\delta)p^{\delta}mB(2m, 1-\delta), \\ \widetilde{\beta_3} = (2/3)(\mu - \sigma/\delta) + (\sigma/\delta)p^{\delta}mB(1.5m, 1-\delta), \end{cases} \tag{3.8}$$

得到

$$\begin{cases} y_1 = \widetilde{\beta_1} = \mu + \sigma t_1, \\ y_2 = 2\widetilde{\beta_2} = \mu + \sigma t_2, \\ y_3 = 1.5\widetilde{\beta_3} = \mu + \sigma t_3, \end{cases} \tag{3.9}$$

其中

$$\begin{aligned} t_1 &= [mp^{\widehat{\delta}}B(m, 1-\widehat{\delta}) - 1]/\widehat{\delta}, \\ t_2 &= [2mp^{\widehat{\delta}}B(2m, 1-\widehat{\delta}) - 1]/\widehat{\delta}, \\ t_3 &= [1.5mp^{\widehat{\delta}}B(1.5m, 1-\widehat{\delta}) - 1]/\widehat{\delta}. \end{aligned}$$

视 (3.9) 式是关于 μ 和 σ 的线性模型，利用最小二乘法，得到 μ 和 σ 的最小二乘估计

$$\widehat{\mu} = \overline{y} - \widehat{\sigma}\overline{t}, \qquad \widehat{\sigma} = \frac{\sum\limits_{i=1}^{3}(t_i - \overline{t})(y_i - \overline{y})}{\sum\limits_{i=1}^{3}(t_i - \overline{t})^2}, \tag{3.10}$$

其中 $\overline{t} = (t_1 + t_2 + t_3)/3$, $\overline{y} = (y_1 + y_2 + y_3)/3$.

在 (3.7) 式中，取 $m = 2$, 可导出 $\widehat{\delta}$ 的解析表达式

$$\widehat{\delta} = \frac{21\widetilde{\beta_3} - 16\widetilde{\beta_2} - 6\widetilde{\beta_1}}{3\widetilde{\beta_3} - 4\widetilde{\beta_2}}, \tag{3.11}$$

对应的

$$\begin{aligned} t_1 &= \frac{1}{\widehat{\delta}}\Big(\frac{2p^{\widehat{\delta}}}{(2-\widehat{\delta})(1-\widehat{\delta})} - 1\Big), \qquad t_2 = \frac{1}{\widehat{\delta}}\Big(\frac{24p^{\widehat{\delta}}}{(4-\widehat{\delta})(3-\widehat{\delta})(2-\widehat{\delta})(1-\widehat{\delta})} - 1\Big), \\ t_3 &= \frac{1}{\widehat{\delta}}\Big(\frac{6p^{\widehat{\delta}}}{(3-\widehat{\delta})(2-\widehat{\delta})(1-\widehat{\delta})} - 1\Big). \end{aligned}$$

进一步，可得到 $\widehat{\mu}$ 和 $\widehat{\sigma}$ 的解析表达式.

对其他已知的 m, 在一定条件下, 可用二分法得到估计方程 (3.7) 中 δ 在区间在 (0,1) 内的数值解.

3.3 估计方程解的存在性

定理 3.1 当 $\frac{\ln 2}{\ln 1.5} < \frac{2\widetilde{\beta}_2-\widetilde{\beta}_1}{1.5\widetilde{\beta}_3-\widetilde{\beta}_1} < 2$ 时，存在偶数 $m=2k$, 使得方程 (3.7) 中的 δ 在 (0,1) 区间内有解.

证 设 $m=2k$, 将方程 (3.7) 的右端记成 $G_k(\delta)$, 即

$$G_k(\delta)=\frac{2B(2m,1-\delta)-B(m,1-\delta)}{1.5B(1.5m,1-\delta)-B(m,1-\delta)},$$

易得

$$G_k(\delta)=\frac{[2\Gamma(4k)\Gamma(2k+1-\delta)-\Gamma(2k)\Gamma(4k+1-\delta)]\Gamma(3k+1-\delta)}{[1.5\Gamma(3k)\Gamma(2k+1-\delta)-\Gamma(2k)\Gamma(3k+1-\delta)]\Gamma(4k+1-\delta)},$$

显然， $G_k(\delta)$ 是 δ 的连续函数. 由

$$G_k(1)=\frac{[2\Gamma(4k)\Gamma(2k)-\Gamma(2k)\Gamma(4k)]\Gamma(3k)}{[1.5\Gamma(3k)\Gamma(2k)-\Gamma(2k)\Gamma(3k)]\Gamma(4k)}=2,$$

$$\lim_{\delta\to 0^+}G_k(\delta)=\lim_{\delta\to 0^+}\frac{\Gamma(3k+1-\delta)}{\Gamma(4k+1-\delta)}\cdot\frac{2\Gamma(4k)\Gamma(2k+1-\delta)-\Gamma(2k)\Gamma(4k+1-\delta)}{1.5\Gamma(3k)\Gamma(2k+1-\delta)-\Gamma(2k)\Gamma(3k+1-\delta)},$$

而 $\lim\limits_{\delta\to 0^+}\frac{\Gamma(3k+1-\delta)}{\Gamma(4k+1-\delta)}=\frac{(3k)!}{(4k)!}$, 利用罗比达法则和特殊函数 $\varphi(\alpha)=\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$, 得

$$\lim_{\delta\to 0^+}\frac{2\Gamma(4k)\Gamma(2k+1-\delta)-\Gamma(2k)\Gamma(4k+1-\delta)}{1.5\Gamma(3k)\Gamma(2k+1-\delta)-\Gamma(2k)\Gamma(3k+1-\delta)}=\frac{4k\Gamma(2k)\Gamma(4k)[\varphi(4k+1)-\varphi(2k+1)]}{3k\Gamma(2k)\Gamma(3k)[\varphi(3k+1)-\varphi(2k+1)]}.$$

进一步，利用 $\varphi(n+1)=\sum\limits_{r=1}^{n}r^{-1}-\gamma$, 其中 γ 为 Euler 常数 (参见 [13], 11 页), 得

$$\lim_{\delta\to 0^+}G_k(\delta)=\frac{\sum\limits_{r=2k+1}^{4k}r^{-1}}{\sum\limits_{r=2k+1}^{3k}r^{-1}}.$$

利用

$$\lim_{k\to\infty}\sum_{r=2k+1}^{4k}r^{-1}=\ln 2,\qquad \lim_{k\to\infty}\sum_{r=2k+1}^{3k}r^{-1}=\ln 1.5$$

及连续函数的介值定理，得证方程 (3.7) 在 $0<\delta<1$ 内必有解.

4 二项 - 广义 Pareto 复合模型的检验

4.1 AD 检验

设 $x_1,x_2,\cdots,x_n$ 是抽自总体 $F(x)$ 的随机样本， $x_{(1)},x_{(2)},\cdots,x_{(n)}$ 为次序样本，

$F_n(x)$ 为样本的经验分布函数，则

$$F_n(x)=\begin{cases}0, & x<x_{(1)},\\ i/n, & x_{(i)}\le x<x_{(i+1)}, \qquad i=1,2,\cdots,n-1,\\ 1, & x_{(n)}\le x.\end{cases}$$

AD 检验法的原理就是构造 AD 统计量. 记

$$A_n^2=n\int_{-\infty}^{\infty}[F_n(x)-F(x)]^2g_1(x)\mathrm{d}F(x), \tag{4.1}$$

其中 $g_1(x)=[F(x)\cdot(1-F(x))]^{-1}$, 以及

$$U_n^2=n\int_{-\infty}^{\infty}[F_n(x)-F(x)]^2g_2(x)\mathrm{d}F(x), \tag{4.2}$$

其中 $g_2(x)=[1-F(x)]^{-1}$.

实际计算时，利用经验分布函数 $F_n(x_{(i)})=i/n$, 代入到 (4.1) 式和 (4.2) 式，可得如下两个常用的 AD[14] 检验统计量

$$A_n^2=-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln Z_{(i)}+\ln(1-Z_{(n+1-i)})]; \tag{4.3}$$

Modified Anderson-Darling(MAD)[12] 检验统计量

$$U_n^2=\frac{n}{2}-2\sum_{i=1}^{n}Z_{(i)}-\sum_{i=1}^{n}\Big[2-\frac{(2i-1)}{n}\Big]\ln(1-Z_{(i)}), \tag{4.4}$$

其中 $Z_{(i)}=\widehat{F}(x_{(i)})$ ($\widehat{F}$ 是将 F 中各个参数由其估计替代后所得到的函数), x_i 为 GPD 样本的第 i 个次序样本.

4.2 KS 检验

Kolmogorov-Smirnov 于 1933 年 [14] 提出的检验法如下.

设原假设 $H_0:\ F(x)=F_0(x)$ ($F_0(x)$ 为某已知分布), 引进统计量

$$D_n^+=\sup_x\{F_n(x)-F(x)\}, \qquad D_n^-=\sup_x\{F(x)-F_n(x)\}, \tag{4.5}$$

令 $D_n=\sup\limits_x|F_n(x)-F(x)|=\max\{D_n^+,\ D_n^-\}$.

实际计算中，利用 $F_n(x_{(i)})=i/n$, 代入到 (4.5) 式，得

$$D_n^+=\max_{1\le j\le n}\Big\{\frac{j}{n}-F(x_{(j)})\Big\}, \quad D_n^-=\max_{1\le j\le n}\Big\{F(x_{(j)})-\frac{j-1}{n}\Big\}, \quad D_n=\max\{D_n^+,D_n^-\}. \tag{4.6}$$

KS 检验方法如下：通过样本 $x_1,x_2,\cdots,x_n$ 计算 D_n, 如果 $D_n<D_0$, 则接受原假设; 否则，拒绝原假设，其中 D_0 是 D_n 在显著度水平 α 下的临界值，即 $P\{D_n>D_0\}=$

$\alpha,\ \alpha\in(0,1)$. D_0 可用查表 [12] 或模拟计算得到. 对于 AD 检验, 方法与 KS 检验类似, 功效一般高于 KS 检验.

GPD 常用于对分布的尾部数据进行建模, 此时其分位数的选取非常重要. 拟合优度检验有多种方法, 这里用最常用的 KS 统计量模拟分位数 D_0, 进而实现对二项 — 广义 Pareto 复合分布模型的拟合优度检验.

5 模拟研究

取样本容量 $n=100,\ m=2,\ p=0.1,0.3,0.5$, 位置参数 $\mu=1$, 刻度参数 $\sigma=0.2$, 形状参数 $\delta=0.2,0.4,0.6,0.8,0.9$, 进行 3000 次 Monte-Carlo 模拟, 得到下面的表 1–3.

表 1 二项 - 广义 Pareto 复合分布的参数加权矩估计模拟 ($p=0.1$)

δ			$\mu=1$		$\sigma=0.2$	
真值	估计偏差	均方误差	估计偏差	均方误差	估计偏差	均方误差
0.2	−0.266816	0.088410	0.030749	0.002342	−0.017891	0.001621
0.4	−0.295317	0.118026	0.043744	0.003519	−0.001519	0.002621
0.6	−0.321331	0.147229	0.058334	0.005354	0.017587	0.004854
0.8	−0.359338	0.177044	0.078901	0.014498	0.041968	0.023770
0.9	−0.392297	0.207124	0.093741	0.036244	0.059767	0.035627

表 2 二项 - 广义 Pareto 复合分布的参数加权矩估计模拟 ($p=0.3$)

δ			$\mu=1$		$\sigma=0.2$	
真值	估计偏差	均方误差	估计偏差	均方误差	估计偏差	均方误差
0.2	−0.273387	0.092502	0.028382	0.001267	0.040964	0.002161
0.4	−0.234464	0.081528	0.027067	0.001219	0.051596	0.003355
0.6	−0.229986	0.085691	0.027750	0.001328	0.060815	0.004808
0.8	−0.248728	0.090941	0.035671	0.002405	0.081374	0.012983
0.9	−0.286564	0.109797	0.043854	0.008792	0.108079	0.228471

表 3 二项 - 广义 Pareto 复合分布的参数加权矩估计模拟 ($p=0.5$)

δ			$\mu=1$		$\sigma=0.2$	
真值	估计偏差	均方误差	估计偏差	均方误差	估计偏差	均方误差
0.2	−0.273744	0.093149	−0.000478	0.000222	0.077053	0.006634
0.4	−0.211351	0.067589	−0.002742	0.000236	0.074057	0.006425
0.6	−0.196344	0.066014	−0.004512	0.000262	0.078248	0.007474
0.8	−0.235821	0.080625	−0.003016	0.000264	0.098965	0.026921
0.9	−0.266233	0.095025	−0.002228	0.000279	0.124800	0.116784

由表 1–3 可知, 当 p 固定, 且 δ 在 0.4 和 0.6 时相对效果最好; 而当其他参数固定, $p=0.5$ 时效果最好. 由表 1–3, 参数 μ 的估计偏差和均方误差都很好, 而参数 σ 的估计

偏差和均方误差仅在 $\delta = 0.9$ 时效果不好，其余结果较好．但参数 δ 的估计偏差和均方误差都不理想．

下面利用 $B_p(\delta) = \widehat{\delta} - \delta$ 修偏，令 $B_p(\delta) = a + bp$，分别取 $p = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7, 0.8,\ 0.9$，位置参数 $\mu = 1$，刻度参数 $\sigma = 0.2$，形状参数 $\delta = 0.4$ 时，进行 3000 次 Monte-Carlo 模拟，得到表 4.

表 4　二项 - 广义 Pareto 复合分布的参数加权矩估计模拟

	$\delta = 0.4$		$\mu = 1$		$\sigma = 0.2$	
p 值	估计偏差	均方误差	估计偏差	均方误差	估计偏差	均方误差
0.1	−0.295317	0.118026	0.043744	0.003519	−0.001519	0.002621
0.2	−0.253847	0.089818	0.040791	0.002374	0.031542	0.002015
0.3	−0.234464	0.081528	0.027067	0.001219	0.051596	0.003355
0.4	−0.224427	0.075144	0.012002	0.000501	0.063953	0.004801
0.5	−0.211351	0.067589	−0.002742	0.000236	0.074057	0.006425
0.6	−0.213781	0.071087	−0.017879	0.000469	0.083451	0.008253
0.7	−0.201791	0.063527	−0.031754	0.001140	0.090751	0.009852
0.8	−0.20607	0.064723	−0.045422	0.002245	0.096864	0.011454
0.9	−0.201468	0.064013	−0.058873	0.003781	0.102676	0.013226

通过 SPSS 软件进行检验得知，$B_p(\delta)$ 与 p 有近似的线性关系，用最小二乘估计得到

$$B_p(\delta) = -0.277 + 0.0991p.$$

对 δ 修偏后得到新的估计偏差和均方误差见表 5–7.

表 5　二项 - 广义 Pareto 复合分布的参数加权矩估计模拟 ($p = 0.1$)

	δ		$\mu = 1$		$\sigma = 0.2$	
真值	估计偏差	均方误差	估计偏差	均方误差	估计偏差	均方误差
0.2	0.000274	0.000162	0.030749	0.002342	−0.017891	0.001621
0.4	0.028227	0.001099	0.043744	0.003519	−0.001519	0.002621
0.6	0.054241	0.003374	0.058334	0.005354	0.017587	0.004854
0.8	0.072248	0.005691	0.078901	0.014498	0.041968	0.023770
0.9	0.095207	0.009587	0.093741	0.036244	0.059767	0.035627

表 6　二项 - 广义 Pareto 复合分布的参数加权矩估计模拟 ($p = 0.3$)

	δ		$\mu = 1$		$\sigma = 0.2$	
真值	估计偏差	均方误差	估计偏差	均方误差	估计偏差	均方误差
0.2	0.026117	0.000856	0.028382	0.001267	0.040964	0.002161
0.4	0.012806	0.000426	0.027067	0.001219	0.051596	0.003355
0.6	0.017284	0.000621	0.027750	0.001328	0.060815	0.004808
0.8	0.001458	0.000288	0.035671	0.002405	0.081374	0.012983
0.9	0.039294	0.001816	0.043854	0.008792	0.108079	0.228471

表 7 二项 - 广义 Pareto 复合分布的参数加权矩估计模拟 ($p = 0.5$)

δ			$\mu = 1$		$\sigma = 0.2$	
真值	估计偏差	均方误差	估计偏差	均方误差	估计偏差	均方误差
0.2	0.046294	0.003932	−0.000478	0.000222	0.077053	0.006634
0.4	0.016099	0.000484	−0.002742	0.000236	0.074057	0.006425
0.6	0.031106	0.001237	−0.004512	0.000262	0.078248	0.007474
0.8	0.008371	0.00031	−0.003016	0.000264	0.098965	0.026921
0.9	0.038783	0.001741	−0.002228	0.000279	0.124800	0.116784

由表 5–7 可以看出，用本文方法对位置参数 μ, 刻度参数 σ 和形状参数 δ 进行估计时都具有较高精度.

在 KS 检验过程中，需要比较 KS 检验统计量 D_n 与备择分布簇在显著度水平 α 下的临界值 D_0 的大小. 本文通过 Monte Carlo 仿真编制二项 – 广义 Pareto 复合分布的临界值表. 以样本容量 $n = 10, 20, 30, 50, 100, 200,\ m = 2,\ p = 0.2$, 位置参数 $\mu = 1$, 刻度参数 $\sigma = 0.1$ 和形状参数 $\delta = 0.2$, 10000 次 Monte Carlo 仿真为例，表 8 给出了样本容量 n 取不同值时 D_n 在显著度水平 $\alpha = 0.01, 0.02, 0.05, 0.10, 0.25, 0.5$ 下的临界值 D_0, 结果见表 8.

表 8 二项 - 广义 Pareto 复合分布的 KS 检验临界值表

n \ α	0.01	0.02	0.05	0.10	0.25	0.5
10	0.9887014	0.9794503	0.9478238	0.8968228	0.8117562	0.6952224
20	0.9906164	0.9813851	0.9514835	0.9227665	0.8596300	0.7759258
30	0.9898212	0.9806506	0.9588009	0.9338348	0.8819270	0.8117564
50	0.9900418	0.9794634	0.9642824	0.9452482	0.9051558	0.8486232
100	0.9920469	0.9856577	0.9736045	0.9596913	0.9303334	0.8899816
200	0.9928557	0.9886292	0.979571	0.970758	0.9494799	0.9209687

如取样本容量 $n = 100,\ m = 2,\ p = 0.2$, 位置参数 $\mu = 1$, 刻度参数 $\sigma = 0.1$ 和形状参数 $\delta = 0.2$, 根据本文方法通过模拟，可得到参数估计分别为 $\widehat{\mu} = 1.028751$, $\widehat{\sigma} = 0.1127270$, $\widehat{\delta} = -0.3018074$, KS 检验统计量 $D_n = 0.87408$ 均小于表 8 中 n=100 时显著度水平 $\alpha = 0.01, 0.02, 0.05, 0.10, 0.25, 0.5$ 下的临界值 D_0, 即该样本服从二项 - 广义 Pareto 复合分布.

参 考 文 献

[1] Bortkiewicz L Von. Variationsbreite and Mittlerer Fehler, Sitzungsber. *Berli. Math. Ges*, 1922, 21: 3–11

[2] Mises R Von. Uber die Variationsbreite Einer Beobachtungsreihe. *Berlin. Math. Ges*, 1923, 22: 3–8

[3] Dodd E L. The Greatest and Least Variate under General Laws of Error. *Trans. Amer. Math. Soc*, 1923, 25: 525–539

[4] Fréhet M. Sur La Loi de Probabilité de 1' Écart Maximum. *Ann. Soc. Polon. Math. Cracovie*, 1927, 6: 93–116

[5] Fisher R A, Tippett L H C. Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *Procs. Cambridge Philos. Soc.*, 1928, 24: 180–190

[6] Gumbel E J. Statistics of Extremes. New York: Columbia University Press, 1958

[7] Balkema A A, de Haan L. Residual Lifetime at Great Age. *Annals of Probability*, 1974, 2: 792–804

[8] Pickands J. Statistical Inference using Extreme Value Order Statistics. *Annals of Statistics*, 1975, 3: 119–131

[9] Feller W. An Introduction to Probability Theory and Its Applications, Volume 1, 3rd Edition. New York: John Willey, 1968

[10] Hosking J M, Wallis J R. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 1987, 29: 339–349

[11] 刘晶. 复合极值分布的参数估计及应用. 天津大学博士论文， 2006
(Liu J. Parameter Estimation and Application for the Compound Extreme Value Distribution. Tianjin University Doctoral Thesis, 2006)

[12] 王松桂，张忠占，程维虎，高旅端. 概率论与数理统计. 北京：科学出版社， 2005
(Wang S G, Zhang Z Z, Cheng W H, Gao L D. Probability Theory and Mathematical Statistics. Beijing: Science Press, 2005)

[13] 森口繁一，宇田川金久，一松信. 数学公式 III. 东京：岩波书店， 2002
(Moriguchi Shigekazu, Udagawa Kanehisa, Ichimatsu Shin. Mathematical Formula III. Tokyo: Iwanami Press, 2002)

[14] 杨振海，程维虎，张军舰. 拟合优度检验. 北京：科学出版社， 2011
(Yang Z H, Cheng W H, Zhang J J. Goodness-of-Fit Tests. Beijing: Science Press, 2011)

Statistical Inference for Binomial-generalized Pareto Compound Extreme Value Distribution Model

ZHANG Xiangyun

(*Tianmu College of Zhejiang A & F University, Lin-an* 311300)

(*E-mail: zlzhangxiangyun@163.com*)

CHENG Weihu

(*College of Applied Sciences, Beijing University of Technology, Beijing* 100124)

(*E-mail: chengweihu@bjut.edu.cn*)

Abstract Extreme value theory is mainly the study on extreme events of small probability & major impact. At present, the compound extreme value distribution has been widely used in hydrology, meteorology, earthquake, insurance, finance and other fields. In this paper, we establish binomial-generalized Pareto compound extreme value distribution model based on extreme value type theorem and PBDH theorem, derive parameter estimation of the established compound model by probability weighted moments, get critical values of Kolmogorov-Smirnov test statistic.

Key words generalized Pareto distribution; Binomial distribution; probability weighted moment; Anderson-Darling test; Kolmogorov-Smirnov test; order statistic; random simulation

MR(2000) Subject Classification 62F07; 62F12

Chinese Library Classification O213.2