# 基于右截尾数据的近极值事件的态密度估计

朱建国[1], 陈 果[2], 黄 超[2]

(1. 南京工业职业技术学院数理学院, 江苏 南京 210046)
(2 东南大学数学系, 江苏 南京 210096)

[摘要] 右截尾数据在实际数据中经常出现, 如材料的疲劳试验等. 本文研究基于右截尾数据的近极值事件的态密度(DOS)估计问题. 首先定义右截尾数据类型下的态密度, 接着推导了平均态密度的精确表达式. 然后在大样本情形, 利用蒙特卡洛(Monte-Carlo)方法、核密度估计方法等来估计平均态密度. 最后应用 Monte-Carlo模拟和亚马尔半岛的夏天气温数据说明了本文的方法, 结果表明平均态密度的近似结果与经验结果均拟合得很好.

[关键词] 右截尾数据, 态密度, 极值统计, 吸引场, 蒙特卡洛方法, 核密度估计
[中图分类号] O212 [文献标识码] A [文章编号] 1001-4616(2010)04-0033-06

# Estimating Density of Near-Extreme Events for Right-Censored Data

Zhu Jianguo[1], Chen Guo[2], Huang Chao[2]

(1. Department of Mathematics, Nanjing Institute of Industry Technology, Nanjing 210046, China)
(2. Department of Mathematics, Southeast University, Nanjing 210096, China)

**Abstract** Right-censored data often occur in the actual data, such as the fatigue test and so on. This paper studies the estimation of density of states(DOS) for right-censored data. We first redefine the DOS and the exact expression of mean DOS is derived. We show the approximate forms of the mean DOS by the theory of extreme value statistics and domains of attraction under the large-sample cases. Finally, Some Monte Carlo simulations and Yamal Peninsula summer temperature data are been used to study the density of near-extreme events, which are found to agree with the theoretical results.

**Key words** right-censored data; density of states; extreme value statistics; domain of attraction; Monte-Carlo method; kernel density estimation

极值统计理论的历史可以追溯到 1928年, Fisher和 Tippett[1]共同提出了 3个著名的极值分布族; von Mises[2]年提出了广义极值分布(GEV), 用一个统一的参数模型来描述 3大极值分布; Pickands[3]提出了广义 Pareto分布模型(GPD), 该模型对于峰值超过阈值方法(POT)的应用有深远的影响; de Haan和 Ferreira[4]进一步给出了极值统计中的吸引场理论.

与极值统计相关的问题是近极值事件(即接近极值的事件)的聚集现象, 即对于一个给定的点近极值事件在总样本中所占的比例大小. 这种近极值事件的聚集现象出现在很多实际问题中, 例如对于保险公司, 保障自己抵御过分大的索赔是非常重要的, 但是同样重要或许更重要的是保障自己抵御众多数目的相当大索赔同时发生. 所以在诸如此类的情况下, 研究极值附近事件的聚集是非常有必要的. 接近极大值的观测数目的渐近表现已经被一些作者研究过[5-7]. 最近的许多文献研究近极值事件的密度来反映近极值事件的聚集. 文[8]提出在独立同分布情况下通过计算近极值事件的态密度(DOS)来分析近极值事件的聚集现象, 并利用极值统计中的理论给出了平均态密度的一些渐近性质以及很好的近似形式, 大大简化了计算; 在此基础上, 文[9]提出了广义态密度(GDOS), 利用过去的极值来估计未来的近极值事件密

收稿日期: 2010-06-07.
通讯联系人: 朱建国, 副教授, 研究方向: 数理统计. E-mail: zhjg@niit.edu.cn

度,是对 DOS的一种推广.

已有若干文献研究普通独立同分布数据的态密度估计问题,但对于截尾数据的态密度研究目前还相对缺乏. 截尾数据又分为左截尾和右截尾,本文主要考虑右截尾数据的态密度估计问题. 所谓右截尾数据,是指对于事先给定的值 $r_0$,若观测值小于 $r_0$就用观测值表示,若大于等于 $r_0$就用 $r_0$表示. 右截尾数据在实际数据中经常出现,如材料的疲劳试验等等. 在本文中,一些常用的统计方法被用来推导平均态密度的估计,如蒙特卡洛(Monte-Carlo)方法、核密度估计方法等. 近似表达式和相关的估计在模拟算例和实际数据中与精确的平均态密度表达式拟合得很好. 第一节给出了右截尾数据的态密度定义并推导得出了平均态密度的精确表达式;进一步地还给出了右截尾数据的平均 态密度的近似和估计形式;第二节对于不同的吸引场给出了一些数值模拟. 第三节通过 1个在实际例子,说明了态密度在实际数据中的应用.

## 1 平均态密度的估计

设 $X_1, X_2, \cdots, X_n$ 为来自分布为 $F$ 的总体中的独立同分布样本,

$$X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}.$$

如果存在数列 $a_n > 0$及 $b_n$ 使得 $a_nX_{(n)} + b_n$依分布收敛于 $G(x)$,则称 $G(x)$为一个极大值分布. 上述分布 $F$称为底分布. 文[1]不加详细的数学证明给出了极值理论中的一个重要结论:如果一个非退化的 $G$存在,则它必是以下 3种类型之一:

$$G(x) = \exp(-e^{-x}),\ \text{所有 } x \tag{1}$$

$$G(x) = \begin{cases} 0, & x < 0, \\ \exp(-x^{-\alpha}), & x > 0, \end{cases} \tag{2}$$

$$G(x) = \begin{cases} \exp(-|x|^{\alpha}), & x < 0, \\ 1, & x > 0 \end{cases} \tag{3}$$

(1)叫做 Gumbel型,(2)叫做 Frechet型,(3)叫做 Weibull型. 在(2)和(3)中,$\alpha > 0$

文[2]提出了广义极值分布(GEV),用一个统一的参数模型来描述三大极值分布. 样本极大值的 GEV分布如下:

$$G(x) = \exp\left\{-\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\},\quad 1 + \xi\frac{x-\mu}{\sigma} > 0, \tag{4}$$

其中 $\mu$是位置参数,$\sigma > 0$是尺度参数,$\xi$是形状参数. 当 $\xi \to 0$对应 Gumbel分布,$\xi > 0$且 $\alpha = 1/\xi$时对应 Frechet分布,$\xi < 0$且 $\alpha = -1/\xi$对应 Weibull分布.

设 $X_1, X_2, \cdots, X_n$为抽自分布为 $F$的总体中的独立同分布样本,若存在常数列 $a_n > 0$及 $b_n$,使得当 $n \to \infty$时,$a_nX_{(n)} + b_n$依分布收敛于(4)型极值分布,则称底分布 $F$属于(4)型分布的吸引场. 文[4]讨论了底分布 $F$属于极值分布 $G$的必要和充分条件;文[9]进一步给出了对不同的极值行分布、数列 $a_n > 0$及 $b_n$的表达式. 这些表达式在推导态密度公式时非常重要.

下面讨论截尾数据的态密度估计问题. 令 $X \sim F(X)$,截断点为 $r_0$,设右截尾数据为 $Z$. 则

$$Z = \begin{cases} X, & X < r_0, \\ r_0, & X \geqslant r_0. \end{cases} \tag{5}$$

因此,$Z$的分布函数为

$$F_Z(z) = \begin{cases} F(z)F(r_0), & z < r_0, \\ 1, & z \geqslant r_0. \end{cases} \tag{6}$$

对于截断点 $r_0$,我们可以得到一列独立同分布截尾数据 $Z_1, Z_2, \cdots, Z_n$. 对于极大值 $Z_{(n)}$,由于数据右截尾,最大值为 $r_0$,所以有可能存在多个 $Z_i = Z_{(n)}$,我们仍只去除 $Z_{(n)}$. 类似于文[8, 9],对于右截尾数据 $Z$,我们定义态密度如下:

$$\rho(r, n) = \frac{1}{n}\left(\sum_{\{Z_i \neq Z_{(n)}\}}^{n-m} \delta[r - (Z_{(n)} - Z_i)] + \sum_{\{Z_i = Z_{(n)}\}}^{n-1} \delta[r - (Z_{(n)} - Z_i)]\right), \tag{7}$$

其中 $Z_1, Z_2, \cdots, Z_n$ $i.i.d \sim F_Z$, $Z_{(n)} \triangleq \max(Z_1, Z_2, \cdots, Z_n)$, $r_i = Z_{(n)} - Z_i$, $i = 1, 2, \cdots, n$, $Z_i \neq Z_{(n)}$, $\delta(\cdot)$是单位脉冲函数.

上述定义的 $\rho(r, n)$ 仍然是一个随机变量, 并不是严格意义上的密度函数, 需要求出右截尾数据的平均态密度 $\overline{\rho(r, n)}$. 因为 $Z_i$, $i = 1, 2, \cdots, n$ 独立同分布, 故省略下标仅考虑其中某一个.

首先, 当 $Z_i \neq Z_{(n)}$ 时,

$$E(\delta[r - (Z_{(n)} - Z)]) = E(E(\delta[r - (Z_{(n)} - Z)] \mid Z_{(n)} = y)) =$$

$$\int_{-\infty}^{r} E(\delta[r - (Z_{(n)} - Z)] \mid Z_{(n)} = y)\, \mathrm{d}F_{Z_{(n)}}(y) + E(\delta[r - (Z_{(n)} - Z)] \mid Z_{(n)} = r_0) P(Z_{(n)} = r_0) =$$

$$\int_{-\infty}^{r} \int_{-\infty}^{\infty} \delta[z - (y - r)]\, \mathrm{d}F(Z \mid Z_{(n)} = y)\, \mathrm{d}F_{Z_{(n)}}(y) + P(Z_{(n)} = r_0)$$

$$\left( \int_{-\infty}^{r} \left( \delta[z - (r_0 - r)]\, \mathrm{d}F(Z \mid Z_{(n)} = r_0) + \delta[r_0 - (r_0 - r)] P(Z = r_0 \mid Z_{(n)} = r_0) \right) \right) =$$

$$\int_{-\infty}^{r} \frac{\int_{-\infty}^{\infty} \delta[z - (y - r)] f(z) F(r_0)\, \mathrm{d}z}{F(y) F(r_0)}\, \mathrm{d}F_{Z_{(n)}}(y) + \int_{-\infty}^{r} \delta[z - (r_0 - r)] f(z) F(r_0)\, \mathrm{d}z P(Z_{(n)} = r_0) \overset{(A)}{=}$$

$$\int_{-\infty}^{r} \frac{f(y - r)}{F(y)}\, \mathrm{d}z F_{Z_{(n)}}(y) + f(r_0 - r) F(r_0) P(Z_{(n)} = r_0). \tag{8}$$

其中 (A) 步用到了狄拉克 $\delta$ 函数的性质.

其次, 当 $Z_i = Z_{(n)}$ 时,

$$E(\delta[r - (Z_{(n)} - Z)]) = E(E(\delta[r - (Z_{(n)} - Z)] \mid Z_{(n)} = y)) =$$

$$\int_{-\infty}^{r} E(\delta[r - (Z_{(n)} - Z)] \mid Z_{(n)} = y) \cdot \mathrm{d}F_{Z_{(n)}}(y) + E(\delta[r - (Z_{(n)} - Z)] \mid Z_{(n)} = r_0) P(Z_{(n)} = r_0) =$$

$$\int_{-\infty}^{r_0} \delta(r)\, \mathrm{d}F(Z = y \mid Z_{(n)} = y)\, \mathrm{d}F_{Z_{(n)}}(y) +$$

$$\delta(r) P(Z = r_0 \mid Z_{(n)} = r_0) P(Z_{(n)} = r_0) = \delta(r) P(Z = r_0) P_{Z_{(n)}} = r_0). \tag{9}$$

因此, 由式 (8) 及式 (9) 得平均态密度为

$$\overline{\rho(r, n)} = \frac{1}{n}\left[ (n - m)\left( \int_{-\infty}^{0} \frac{f(y - r)}{F(y)}\, \mathrm{d}F_{Z_{(n)}}(y) + f(r_0 - r) F(r_0) P(Z_{(n)} = r_0) + \right.\right.$$

$$\left.\left. (m - 1)\, \delta(r) P(Z = r_0) P(Z_{(n)} = r_0) \right)\right]. \tag{10}$$

对于 $y < r_0$, 我们有 $P(Z \leqslant y) = F(y) F(r_0)$, 所以可以得到

$$P(Z_{(n)} \leqslant y) = (F(y) F(r_0))^n.$$

进一步可得 $P(Z = r_0) = 1 - F^2(r_0)$, $P(Z_{(n)} = r_0) = 1 - F^{2n}(r_0)$. 从而 (10) 化简后可得

$$\overline{\rho(r, n)} = \frac{1}{n}\left[ (n - m)\left( nF^n(r_0) \int_{-\infty}^{r} f(y - r) F^{n-2}(y) f(y)\, \mathrm{d}y + f(r_0 - r) F(r_0) (1 - F^{2n}(r_0)) \right) + \right.$$

$$\left. (m - 1)\, \delta(r) (1 - F^2(r_0)) (1 - F^{2n}(r_0)) \right]. \tag{11}$$

$\overline{\rho(r, n)}$ 的表达式非常复杂, 我们首先在 $n \to \infty$ 时对其进行近似简化. 当 $n \to \infty$ 时, 有 $F^n(r_0) \to 0$, 所以平均态密度的近似形式为

$$\overline{\rho(r, n)} \approx \frac{1}{n}[(n - m) f(r_0 - r) F(r_0) + (m - 1)\, \delta(r) (1 - F^2(r_0))]. \tag{12}$$

在实际情况中观测值的分布一般未知, 我们需要对 $\overline{\rho(r, n)}$ 进行估计:

$$\widehat{\overline{p(r, n)}} = \frac{1}{n}[(n - m) \hat{f}(r_0 - r) \hat{F}(r_0) + (m - 1) \delta(r) (1 - \hat{F}^2(r_0))]. \tag{13}$$

因为 $P(Z < r_0) = F^2(r_0) \approx 1 - \frac{m}{n}$, 所以 $F(r_0)$ 的估计为

$$\hat{F}(r_0) = \sqrt{1 - \frac{m}{n}}. \tag{14}$$

对于密度函数 $f(x)$ $(x < r_0)$, 我们采用核密度估计. 给定一个核函数 $K$ 和一个称为带宽的正数 $h$, 核密度估计定义为

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{Z - Z_i}{h}\right) \tag{15}$$

常用的核函数有 boxcar核、Gaussian核、Epanechnikov核和 tricube核等. 核密度估计对核 $K$ 的选择并不重要,但对带宽 $h$的选择则是至关重要的. 这里我们考虑交叉验证(CV)方法来确定带宽 h 风险的交叉验证估计定义为

$$\hat{J}(h) = \int[\hat{f}(z)]^2\,\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(Z_i),$$

这里 $\hat{f}_{-i}$为在删去第 $i$个观测之后得到的密度估计. 称 $\hat{J}(h)$ 为交叉验证得分或估计的风险.

我们根据使 $\hat{J}(h)$ 最小来确定带宽 $h$. 对于本文需要估计的 $f(x)$, 有 $Z_i < r_0$. 文[10]给出了 $\hat{J}(h)$ 的一个简单表达式:

$$\hat{J}(h) = \frac{1}{hn^2}\sum_i\sum_j K^*\left(\frac{Z_i - Z_j}{h}\right) + \frac{2}{nh}K(0) + O\left(\frac{1}{n^2}\right),$$

这里 $K^* = K^{(2)}(x) - 2K(x)$, 而 $K^{(2)}(x) = \int K(z-y)K(y)\,\mathrm{d}y$. 当 $K$ 为一个 $N(0,1)$ 的 Gaussian核时,那么 $K^{(2)}$ 为正态分布 $N(0,2)$ 的密度函数.

## 2　数值模拟

本节, 我们对于属于 3个不同吸引场的底分布举例来观察平均态密度的近似及估计结果与精确表达式的拟合程度.

例 **1**　$F \in D(G_{\xi>0})$.

取 $F(x)$ 为 Frchet分布:

$$f(x;\gamma) = \gamma x^{-1-\gamma}\exp(-x^{-\gamma}),\quad \gamma > 0,\ x > 0,$$

属于 Frchet类吸引场. 设截断点 $r_0 = 20$, 取 $\gamma = 1$, 从而由式(6)得到实际数据 $Z$ 的分布函数 $F_Z(z)$. 首先我们从分布 $F_Z(z)$ 中产生 5 000个随机数. 计算其中等于 $r_0$ 的个数 $m$, 由式(14)估计 $\hat{F}(r_0)$. 由其中的 $5000-m$ 个小于 $r_0$的数据, 根据 CV方法选择 $h = 0.1468$, 由式(15)得核密度估计 $\hat{f}(z)$. 拟合的结果见图 1.

例 **2**　$F \in D(G_{\xi<0})$.

取 $F(x)$ 为 Beta分布 Beta$(a, b)$:

$$f(x;a,b) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1},\quad 0 < x < 1, \tag{16}$$

属于 Weibull类吸引场. 设截断点 $r_0 = 0.7$, 取 $a = 10$, $b = 10$, 从而由式(6)得到实际数据 $Z$ 的分布函数 $F_Z(z)$. 首先我们从分布 $F_Z(z)$ 中产生 5 000个随机数. 计算其中等于 $r_0$ 的个数 $m$, 由式(14)估计 $\hat{F}(r_0)$. 由其中的 $5000-m$ 个小于 $r_0$的数据, 根据 CV方法选择 $h = 0.0401$, 由式(15)得核密度估计 $\hat{f}(z)$. 拟合的结果见图 2.

例 **3**　$F \in D(G_{\xi=0})$ $(a_n \to \infty)$.

取 $F(x)$ 为 Weibull分布:

$$f(x;\alpha) = \alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}},\quad \alpha,\ \lambda > 0,\ x > 0,$$

属于 Gumbel类吸引场,同时当 $n \to \infty$时归一化参数 $a_n \to \infty$. 设截断点 $r_0 = 10$, 取 $\alpha = \frac{1}{2}$, $\lambda = 1$, 从而由式(6)得到实际数据 $Z$ 的分布函数 $F_Z(z)$. 首先我们从分布 $F_Z(z)$ 中产生 5 000个随机数. 计算其中等于 $r_0$的个数 $m$, 由式(14)估计 $\hat{F}(r_0)$. 由其中的 $5000-m$ 个小于 $r_0$的数据, 由 CV方法选择 $h = 0.0975$, 由式(15)得核密度估计 $\hat{f}(z)$. 拟合的结果见图 3.

例 **4**　$F \in D(G_{\xi=0})$ $(a_n \to 0)$.

取 $F(x)$ 为正态分布 $N(\mu, \sigma^2)$:

$$f(x;\ \mu,\ \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{17}$$

属于 Gumbel类吸引场, 同时当 $n \to \infty$时归一化参数 $a_n \to 0$ 设截断点 $r_0 = 1.5$, 取 $\mu = 0$, $\sigma^2 = 1$, 从而由式 (6) 得到实际数据 $Z$ 的分布函数 $F_Z(z)$. 首先我们从分布 $F_Z(z)$ 中产生 5 000个随机数. 计算其中等于 $r_0$ 的个数 $m$, 由式 (14) 估计 $\hat{F}(r_0)$. 由其中的 $5\,000 - m$ 个小于 $r_0$ 的数据, 根据 CV 方法选择 $h = 0.577\,4$, 由式 (15) 得核密度估计 $\hat{f}(z)$. 拟合的结果见图 4

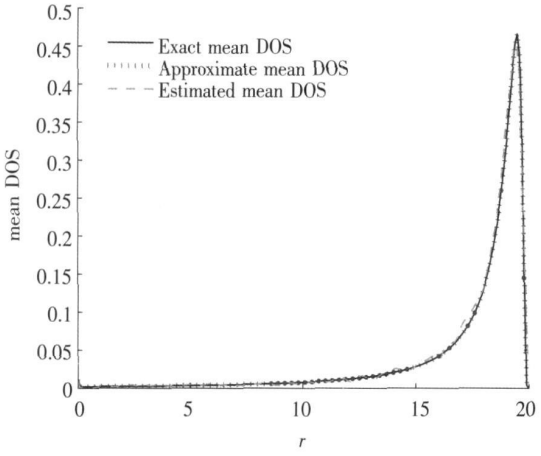


图 1 Fréchet 分布平均态密度的模拟结果

Fig.1 Results of Fréchet distribution mean DOS

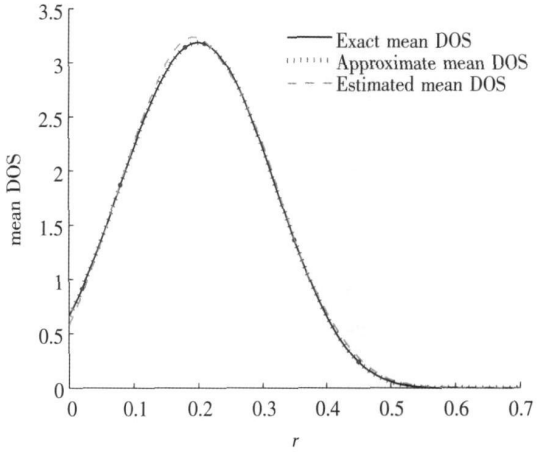


图 2 Beat 分布平均态密度的模拟结果

Fig.2 Results of Beat distribution mean DOS

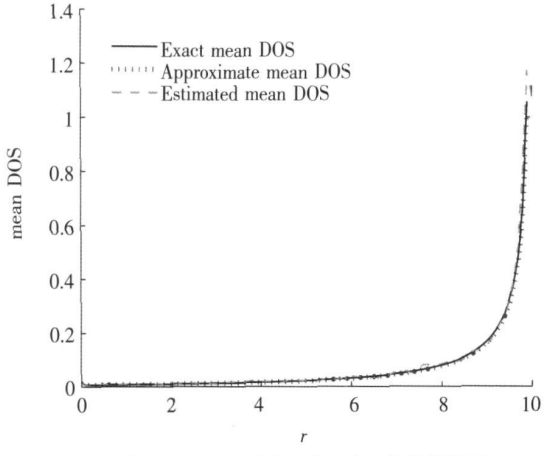


图 3 Weibull 分布平均态密度的模拟结果

Fig.3 Results of Weibull distribution mean DOS

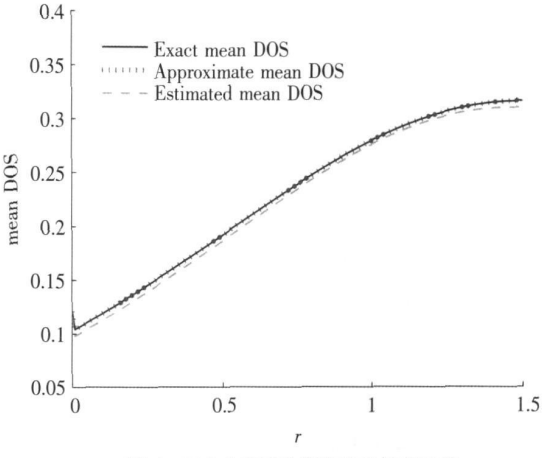


图 4 正态分布平均态密度的模拟结果

Fig.4 Results of Normal distribution mean DOS

例 **5** $F \in D(G_{\xi=0})$ $(a_n \to a)$.

取 $F(x)$ 为指数分布 $\exp(\lambda)$;

$$f(x;\ \lambda) = \lambda e^{-\lambda x},\quad \lambda > 0,\ x > 0,$$

属于 Gumbel类吸引场, 同时当 $n \to \infty$时归一化参数 $a_n$ 趋于某个常数 $a$. 设截断点 $r_0 = 3$, 取 $\lambda = 1$, 从而由式 (6) 得到实际数据 $Z$ 的分布函数 $F_Z(z)$. 首先我们从分布 $F_Z(z)$ 中产生 5 000个随机数. 计算其中等于 $r_0$ 的个数 $m$, 由式 (14) 估计 $\hat{F}(r_0)$. 由其中的 $5\,000 - m$ 个小于 $r_0$ 的数据, 根据 CV 方法选择 $h = 0.081\,2$, 由式 (15) 得核密度估计 $\hat{f}(z)$. 拟合的结果见图 5

## 3 实际例子

这里我们使用亚马尔半岛的夏天气温数据来检验我们上面得到的有关右截尾数据平均态密度的结果. 我们仍将这组数据每 100年分成一组, 共分成 40组. 设截断点 $r_0 = 2$, 然后对于每一组应用式 (7) 计算 $\rho(r,\ n)$. 对所有组求期望我们得到 $\overline{\rho(r,\ n)}$ 的经验结果. 计算其中等于 $r_0$ 的个数 $m$, 由式 (14) 估计

$\hat{F}(r_0)$. 利用其中小于 $r_0$ 的数据, 根据 CV 方法选择 $h = 0.0856$, 由式(15)得核密度估计 $\hat{f}(z)$, 从而得到 $\overline{\rho_2(r, n)}$ 的近似结果. $\overline{\rho(r, n)}$ 的经验结果和近似结果都在图 6中给出. 可以看出两者拟合得很好.

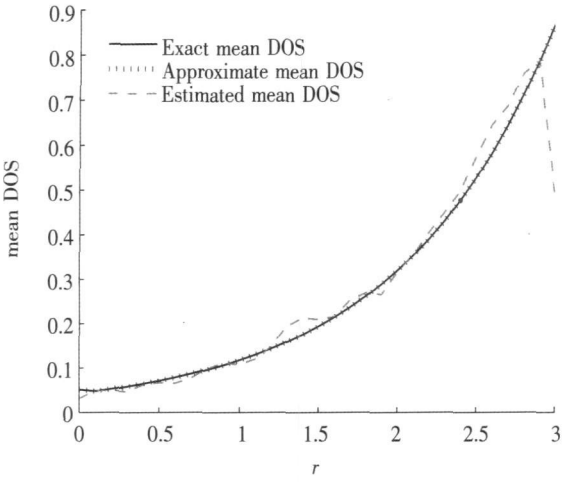


图 5　指数分布平均态密度的模拟结果

Fig.5　Results of Exponential distribution mean DOS

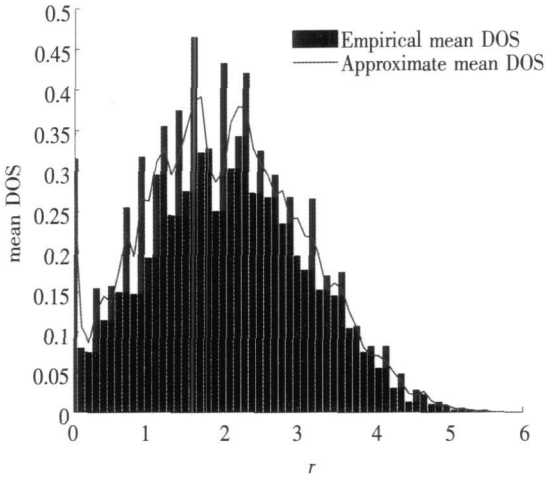


图 6　亚马尔半岛夏天气温数据的模拟结果

Fig.6　Results of Yamal summer temperature data

[参考文献]

[1] Fisher R A, Tippet L H C. Limiting forms of the frequency distributions of the largest or smallest member of a sample[J]. Proceedings of Cambridge Philosophical Society, 1928, 24: 180.

[2] von Mises R. La distribution de la plus grande de $n$ valeurs[J]. Revista Mathematica Union Interbalcanique, 1936, 1: 141–160.

[3] Pickands J. Statistical inference using extreme order statistics[J]. The Annals of Statistics, 1975, 3(1): 119–131.

[4] de Haan L, Ferreira A. Extreme Value Theory: An Introduction[M]. New York: Springer, 2006.

[5] Brands J J A M, Steutel F W, Wilms R J G. On the number of maxima in a discrete sample[J]. Statistics and Probability Letters, 1994, 20: 209–217.

[6] Li Y. A note on the number of records near the maximum[J]. Statistics and Probability Letters, 1997, 43: 153–158.

[7] Pakes A G, Steutel F W. On the number of records near the maximum[J]. The Australian Journal of Statistics, 1997, 39(2): 179–192.

[8] Sanjib Sabhapandit, Satya N, Majumdar. Density of near-extreme events[J]. Physical Review Letters, 2007, 98(14): 41–47.

[9] Lin J G, Huang C, Zhuang Q Y. Estimating generalized state density of near-extreme events and its applications in analyzing stock data[J]. Insurance: Mathematics and Economics, 2010, 47: 13-20.

[10] Wasserman L. All of Nonparametric Statistics[M]. New York: Springer-Verlag, 2007.

[责任编辑: 丁　蓉]