第 41 卷第 6 期 数学的实践与认识 Vol.41, No.6
2011年 3 月 MATHEMATICS IN PRACTICE AND THEORY Mar., 2011

管理科学

重尾分布的尾部指数估计及沪深股市实证分析

刘维奇 [1,2], 赫英迪 [2,3], 陈 琳 [2,4]

(1. 山西大学 管理科学与工程研究所, 山西 太原 030006)
(2. 山西大学 数学科学学院, 山西 太原 030006)
(3. 广东茂名职业技术学院, 广东 茂名 525000)
(4. 山西大学 工程学院, 山西 太原 030013)

摘 要: 重尾分布尾部指数 α 的估计依赖于样本中所用顺序统计量个数 k 的选取. 本文介绍了估计 α 时选择 k 的两类不同的方法: Sum-plot 方法和 Bootstrap 方法, 并对 Hall 提出的 Bootstrap 方法作了改进, 称为 M-Bootstrap 方法. 本文利用上述方法对已知分布进行 Monte-Carlo 模拟, 研究它们的可行性, 然后对上海和深圳两市股指数据进行了实证分析. 计算结果表明, 上海和深圳股指收益率具有重尾性, 是右偏态的, 右尾厚于左尾. 通过几种方法计算的结果比较发现 Sum-plot 方法和 M-Bootstrap 方法在估计重尾指数上精确性较高一些, 而且不受异常值的影响.

关键词: 重尾分布; 重尾指数; Hill 估计; Sum-plot 方法; Bootstrap 方法

1 引言

重尾分布特征在许多领域中普遍存在, 比如经济、金融、通信、水文学、气象学等领域中高频时间序列数据的边际分布几乎都是重尾的. 因此, 估计极值事件发生的概率是十分必要的. 根据统计学的理论, 随机变量的特性应通过随机变量的概率分布描述. 因此, 欲捕捉极端事件发生的概率, 必须能正确描述分布的重尾程度, 即准确估计出重尾分布的尾部指数, 我们称为重尾指数. 所以, 重尾指数估计一直是许多学者关注的问题.

近年来, 学者们给出了估计重尾指数许多方法. 早在 1949 年 Zipf[1] 针对经济以及心理学问题中典型的重尾分布例子利用 Zipf-plot 给出了重尾指数的估计. 随后 Fama & Roll[2]、Press[3] 和 Zolotare[4] 分别研究了稳定分布指数的估计方法. Pickands[5] 给出了基于顺序统计量的 Pickands 估计, 并指出对于随机变量 X, 无论整体上服从何种分布, 如果能给定其充分大的一个门限 u(临界值), 则随机变量 X 超过 u 的条件分布收敛于广义帕累托分布 (GPD). Pickands 估计激起了众多重尾指数估计方法的提出. 如 1975 年 Hill[6] 给出了基于条件极大似然估计的 Hill 估计, 1985 年 Csörgö等 [7] 提出的核估计方法, 随后有 Smith[8−9] 的参数估计方法, Dekkers 等 [10] 的矩估计算法, Kratz & Resnick[11] 建议的基于最小二乘方法的 qq-plot 估

收稿日期: 2008-02-10
资助项目: 教育部人文社会科学研究项目 (07JA630027, 06JA630035); 山西省高校人文社科重点研究基地项目 (20083006)

计, Beirlant 等 [12−13] 利用迭代方法给出的修正 Hill 估计, Feuerverger & Hall[14] 给出的在 Pareto 分布的条件下, 能使估计的偏差减小而不增加方差的两种估计方法, 以及其它一些修正的 Hill 估计等等. 这些估计方法总体上分为参数估计 (Smith) 和非参数估计 (Hill、Csörgö et al、Dekkers et al、Beirlant et al 等). 前者要求随机变量的条件分布收敛于严格的 GPD, 后者仅要求随机变量的条件分布收敛于极值分布. 无论是参数估计还是非参数估计, 这些估计方法都与估计中所用顺序统计量个数 k 有关, k 的选取关系到估计的精确性, k 的偏大或者偏小都会造成估计误差. 所以, 估计重尾指数时究竟选取多少个较大的顺序统计量, 重尾分布的尾部究竟从哪开始, 是估计中首要考虑的问题.

如何选取最优的 k 引起了学者们的广泛关注, 提出了不少选取方法. 其中一类就是作图方法, 比如 Hill 提出的 Hill-plot 图、Kratz & Resnick 提出的 qq-plot 图、Beirlantetal 提出的 Pareto 分位数图和 Resnick & Stăică[15] 给出的 Alt Hill-plot 图等, 这些作图法都有一定的优越性, 但整体而言它们都不能适用于一般情况的重尾分布. 像 Hill-plot 图、qq-plot 图, 当随机变量的分布服从 Pareto 分布时, 这两种方法表现出十分优良的性质 (见图 1). 图形的稳定区域很明显, 起始点 k 值很好寻找. 但是一旦随机变量的分布服从的不是 Pareto 分布, 而是广义 Pareto 分布 (或称 Pareto 型分布) 时, 它们却不能很好地选取 k, 甚至是无法选取 k(见图 2).

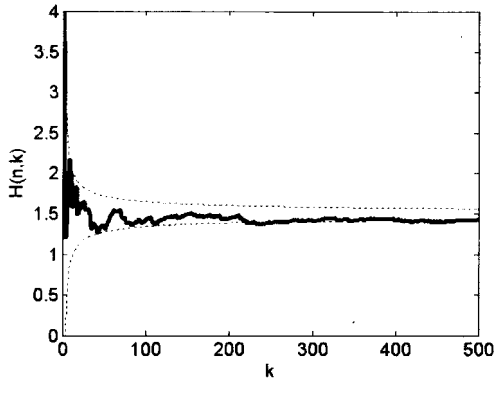

图 1 Pareto 分布的 Hill-plot 图 (n=500)[1]

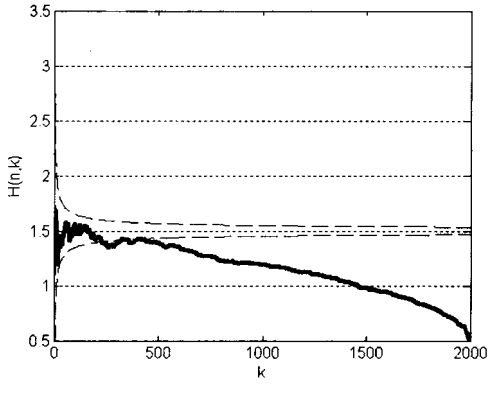

图 2 IGa(1.5,1) 分布的 Hill-plot 图

还有一类方法是以估计极值指数 (重尾指数的倒数) 的均方误差 (MSE) 最小为标准来确定 k. 最优的 k 应该与 MSE 最小相一致, 理论上 MSE(方差与偏差的平方和) 与 k 有关, 增大 k, 方差减小, 偏差增大, 反之, 减小 k 方差增大, 偏差减小. 只有权衡方差与偏差, 使 MSE 最小, 选取的 k 才是最优的. 但是, MSE 还与未知分布尾部指数 α 和二阶参数 ρ 有关, 不能直接应用于实际问题中. 基于此, 我们引入 Sousa 提出的 Sum-plot 方法和 Hall 以及 Danielsson 等提出的 Bootstrap 方法, 我们还对 Hall 提出的 Bootstrap 作了修正, 称为 M-bootstrap 方法. 本文证明使用 Sum-plot 方法和 M-Bootstrap 方法, 能够较好地解决 k 的选择问题, 结合 Hill 估计能够实现对重尾指数的估计.

早在 1963 年 Mandelbrot[16] 在研究棉花价格时就发现, 金融资产收益率不遵循正态分布, 存在尖峰厚尾特征, 呈重尾分布. 许多学者在对股票市场收益率的统计分析表明, 收益率

1 图中虚线表示 $\alpha \pm \frac{k}{\quad}\alpha$, 其中 $\frac{k}{\quad}\alpha$ 表示 Hill 估计的标准偏差, α 是重尾指数

的密度函数的尾部要比正态分布厚, 即较极端情形 (如高损失或高回报) 发生的概率要高于正态分布所表示的概率. 金融领域的极端事件的特点就是发生的概率小, 但后果往往非常严重. 因此有必要准确描述其分布的重尾程度, 这在风险管理中是值得充分重视的问题. 中国的股票市场是否存在重尾现象? 本文对沪深股市股指收益率进行实证分析, 准确估计其重尾指数, 以便更好地度量风险, 管理风险.

本文的结构是先简要介绍极值理论和重尾分布, 然后详细讨论选取 k 的 Sum-plot 法和 M-Bootstrap 法, 为比较方法的稳健性, 分别对已知分布进行了模拟, 重点对沪深两市的股指数据进行实证分析, 并考察异常值对估计的影响, 最后是本文的结论.

2 极值理论与重尾分布

极值理论主要研究随机样本或随机过程中极端事件发生的概率及其统计推断, 是研究分布尾部行为的一个重要工具. 我们用 $X_n^{(1)} \geq X_n^{(2)} \geq \cdots X_n^{(n)}$ 表示分布函数为 F 的独立同分布随机变量 $X_1, X_2, \cdots X_n$ 的顺序统计量, 已经证明存在数列 $a_n > 0$ 和 b_n, 当 n 趋于无穷大时, 使标准化后的最大观测值 $(X_n^{(1)} - b_n)/a_n$ 弱收敛于一个非退化的分布函数 G_γ, 即 $\lim\limits_{n\to\infty} p\left(\frac{X_n^{(1)}-b_n}{a_n} \leq x\right) = G_\gamma(x)$, 而且 G_γ 必属于下列极值分布中的其中一类 ($\gamma > 0, \gamma = 0, \gamma < 0$).

$$G_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), & \gamma \neq 0 \\ \exp(-\mathrm{e}^{-x}), & \gamma = 0 \end{cases} \tag{1}$$

称参数 γ 为极值指数, 它表示极端事件出现的频率, 刻画随机变量 X 尾部的形状, γ 越大, 分布的尾部越厚. 对于 γ >0, 记 γ=1/ α, 随机变量 X 的分布属于极值分布的 Pareto 型分布, 也称极值二型分布或 Fréchet 分布, 记作 $F \in MDA(\Phi_\alpha)$, 其中 $\Phi_\alpha = \exp(-x^{-\alpha})$, $x > 0$, α >0.

$F \in MDA(\Phi_\alpha)$ 等价于 $\overline{F}(x) = 1 - F(x)$ 在无穷远处以指数 $-\alpha$ 正则变化, 即当 $x \to \infty$

$$\overline{F}(x) = 1 - F(x) = x^{-\alpha} L(x) \tag{2}$$

其中 L(x) 是慢变化函数, 即满足 $\lim\limits_{t\to\infty} L(tx)/L(t) = 1$, $x > 0$. 此即分布函数 $F(x)$ 满足一阶正则变化性质. 如果存在无穷远处不改变符号函数 $A(t)$, 使得对 $x > 0$,

$$\lim_{t\to\infty} \left(\left(\frac{1-F(x)}{1-F(t)} - x^{-\alpha} \right) / A(t) \right) = x^{-\alpha} \frac{x^{\alpha\rho} - 1}{\alpha\rho} \tag{3}$$

称分布函数 $F(x)$ 满足二阶正则变化性质, 其中 $\rho \leq 0$ 称为二阶参数.

重尾分布 一般地, 我们称随机变量 X 具有重尾分布 F, 如果 F 满足式 (2), 称 α 为重尾分布 F 的尾部指数或重尾指数, α 越小, 其尾部越厚.

Pareto 分布 称随机变量 X 服从 Pareto 分布, 如果其分布函数 F 满足

$$1 - F(x) = cx^{-\alpha}, \quad x > c, c > 0, \alpha > 0 \tag{4}$$

Hill 估计 如果 F 满足式 (2), 称

$$H_{n,k}^{-1} = \frac{1}{k} \sum_{i=1}^{k} \log X_n^{(i)} - \log X_n^{(k+1)}, \quad 1 \leq k < n \tag{5}$$

为重尾指数 α 的 Hill 估计, k 是估计中所使用较大顺序统计量的个数.

由此可见, Hill 估计与所使用较大顺序统计量的个数 k 紧密相关, k 一旦被确定, 重尾指数自然就能够算出.

3 重尾指数的估计

重尾指数估计的关键在于确定使用顺序统计量个数 k. 一旦正确估计出 k, 应用 Hill 估计不难得到重尾指数的估计. 我们引入 Sousa 提出的 Sum-plot 方法和 Danielsson 等提出的 Bootstrap 方法, 并受 Danielsson 等的启发, 对 Hall 提出的 Bootstrap 方法作了改进, 我们称为 M-Bootstrap 方法.

3.1 Sum-plot 方法

Sum-plot 方法是于 2002 年 Sousa[17] 在其博士论文中提出的, 该方法是一种通过作图来确定 k 的方法. Sum-plot 方法是基于 $\{(k, S_k), 1 \le k < n\}$ 应该是一条直线的理论依据来选取 k. Sousa 通过对不同分布进行的模拟, 得出无论 $0 < .\alpha < 2$ 还是 $\alpha > 2$, Sum-plot 方法对绝大多数分布而言, 都较其它方法优越. 这里随机变量

$$S_k = \sum_{i=1}^{k} i(\log X_n^{(i)} - \log X_n^{(i+1)}) = \sum_{i=1}^{k} (\log X_n^{(i)} - \log X_n^{(k+1)}), 1 \le k < n \tag{6}$$

如果选择 k 使 $X_n^{(k+1)}$ 足够大, 使其满足式 (2), 那么对任意 $x > X_n^{(k+1)}$, $S_k \sim \alpha^{-1}k$. 近似式表明图形中直线的斜率等于.α^{-1}, 可以通过如下线性回归模型将.α^{-1} 估计出来.

$$S_i = \beta_0 + \beta_i + \varepsilon_i, i = 1, \cdots, k \tag{7}$$

容易发现, 参数.α^{-1} 的估计值等于基于最小二乘法方法 (LSE) 估计的回归模型的斜率 $\widehat{\beta}_1$, 即

$$\widehat{\alpha^{-1}} = \widehat{\beta}_1 = \frac{k}{k-1} H_{n,k}^{-1} - \frac{1}{k-1} \log X_n^{(1)}$$

进一步, 如果 β_0=0, 则 $\hat{\alpha}^{-1} = \widehat{\beta_1} = H_{n,k}^{-1}$, 正好就是 Hill 估计.

3.2 Hall 提出的 Bootstrap 方法

Bootstrap 方法是在 1979 年由 Efron[18] 首先提出的, 该方法通过对样本的经验分布进行随机抽样, 得到 Bootstrap 子样本, 然后再进行统计量的估计. 研究结果表明 Bootstrap 方法具有很好的大样本性质.

1990 年 Hall[19] 将 Bootstrap 方法用于重尾指数估计中确定所使用选取顺序统计量的个数. 该方法是基于极小化极值指数估计的渐近均方误差 (AMSE) 来确定 k. 其中

$$AMSE(n,k) = AsyE(\gamma_n(k) - \gamma)^2, 1 \le k < n \tag{8}$$

确定的 $k = \arg\min_k AsyE(\gamma_n(k) - r)^2$, 这里 $\gamma_n(k) = H_{n,k} = \frac{1}{k}\sum_{i=1}^{k} \log X_n^{(i)} - \log X_n^{(k+1)}$, γ 为极值指数.

已经证明, 当 $k \to \infty$, $k/n \to 0$ 时 $\gamma_n(k)$ 依概率收敛于 γ, 并且当 γ 和 ρ 已知时

$$\sqrt{k}(\gamma_n(k) - \gamma) \xrightarrow{d} N(b, \gamma^2)$$

事实上 k 值平衡了 $E(\gamma_n(k) - \gamma)^2$ 的渐进方差和偏差. 因为 γ 和 ρ 一般未知, 需要从总体样本中抽取子样本 $\mathrm{X}_{n_1}^* = \{x_1^*, x_2^*, \cdots x_{n_1}^*\}$, $n_1 = n^\beta (0 < \beta < 1)$, 称 $\mathrm{X}_{n_1}^*$ 为 Bootstrap 子样本, 用 $X_{n_1}^{(1)} \ge \cdots \ge X_{n_1}^{(n_1)}$ 表示其顺序统计量, 通过极小化

$$\widehat{AMSE}(n_1,k_1) = E((\gamma_{n_1,}^*(k_1) - \gamma_n(k))^2 | \mathrm{X}_n) \tag{9}$$

和关系 $k = k_1(n/n_1)^\mu$ 来确定 k_1 与 k. 事实上预设了 k 与 n 之间的幂指数关系式 $k = cn^\mu$, Hall 建议选取 μ=2/3, β=1/2 Caers & Dyck[20] 通过模拟, 证实 μ=2/3, β=1/2 是最优选择. 该方法依赖于 γ 和 ρ, 并且 AMSE 与 k 有关.

3.3 Danielsson 等提出的 Bootstrap 方法

Danielsson 等 [21] 对 Hall 的方法作了改进, 使用新的统计量来代替 (4) 中的 $\gamma_n(k)$. 引入统计量

$$M_n(k)=\frac{1}{k}\sum_{i=1}^{k}(\log X_n^{(i)}-\log X_n^{(k+1)})^2, 1\le k<n \tag{10}$$

已经证明, 当 $k\to\infty, k/n\to 0$ 时 $M_n(k)/(2\gamma_n(k))$ 依概率收敛于 γ, 统计量 $M_n(k)/(2\gamma_n(k))-\gamma_n(k)$ 和 $\gamma_n(k)-\gamma$ 有相似的渐近性质, 并且在一定条件下极小化 $Asy(M_n(k)-2\gamma_n(k)^2)^2$ 和极小化 $AsyE(\gamma_n(k)-\gamma)^2$ 可以得到同阶量 (相对于 n) 的 k.

因此, 根据 Bootstrap 子样本 $\mathrm{X}_{n_1}^*$, 选用统计量:

$$Q(n_1,k_1)=E((M_{n_1}^*(k_1)-2(\gamma_{n_1}^*(k_1))^2)^2|\mathrm{X}_n) \tag{11}$$

通过最小化 $Q(n_1,k_1)$ 来确定 k_1. 为了确定 k, 还需要另一个 Bootstrap 子样本 $\mathrm{X}_{n_2}^*$, $n_2=n_1^2/n$, 然后利用与确定 k_1 相同程序来确定 k_2. 应用 k、k_1 和 k_2 之间的关系

$$k=\frac{k_1^2}{k_2}\left(\frac{(\log k_1)^2}{(2\log n_1-\log k_1)^2}\right)^{\frac{\log n_1-\log k_1}{\log n_1}} \tag{12}$$

来确定 k.

3.4 改进的 Bootstrap 方法

我们受 Danielsson 等提出的 Bootstrap 方法的启发, 用 $M_n(k)/(2\gamma_n(k))$ 代替 $\gamma_n(k)$, 统计量变为 [22]

$$\widehat{AMSE}_M(n,k)=AsyE(\gamma_n(k)-M_n(k)/2\gamma_n(k))^2, 1\le k<n \tag{13}$$

根据 Bootstrap 子样本 $\mathrm{X}_{n_1}^*$, 通过极小化 $\widehat{AMSE}_M(n_1,k_1)$ 和关系 $k=k_1(n/n_1)^\mu$ 来确定 k_1 与 k.

4 重尾指数的 Monte-Carlo 模拟

为了更好的说明问题, 我们选用三种熟知的重尾分布, Pareto 分布 Pareto (1.5)、t- 分布 t(3) 和逆 Γ- 分布 IGa(α,1), 分别采用 Sum-plot 方法、Danielsson 等提出的 Bootstrap 方法 (为简便, 在本文中称为 Bootstrap 方法) 和改进的 Bootstrap 方法 (称为 M-Bootstrap 方法), 进行模拟. 结果表明, Sum-plot 方法、Bootstrap 方法和 M-Bootstrap 方法都能作为 Hill 估计中选择 k 的有力工具, 它们和 Hill 估计结合起来估计重尾指数将是有效的.

我们首先将 Sum-plot 方法用于 Pareto 分布, 从 Pareto(1.5) 分布分别抽取样本容量 n 为 500、2000 的样本, 模拟结果见图 3.

由图 3 可以看出, Pareto 分布的 Sum-plot 图近似为直线, 当样本容量 n 为 2000 时, 图形几乎就是一条直线. 这说明针对 Pareto 分布, 选择的 $k=n$ 即可. 我们将选择的 k 代入到 Hill 估计中计算出重尾指数分别为:$H_{500,500}=1.5463$, $H_{2000,2000}=1.5263$, 接近真实值 1.5.

接下来我们运用 Sum-plot 方法对 $t(3)$ 分布和 IGa(α,1) 分布模拟, 选取样本容量分别为 1000 和 2000. 首先画出它们全部数据的 Sum-plot 图 (见图 4).

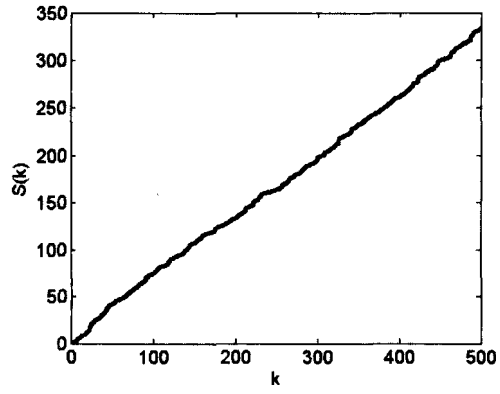

图 3a Pareto (1.5) 分布的 Sum-plot 图 (n=500)

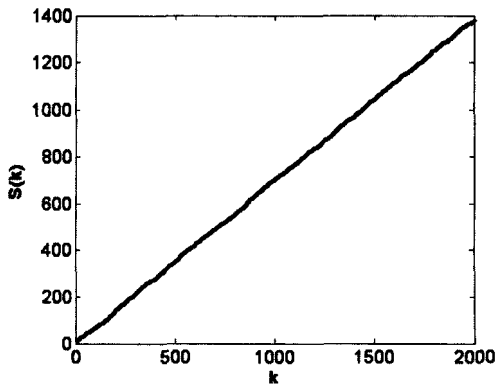

图 3b Pareto (1.5) 分布的 Sum-plot 图 (n=2000)

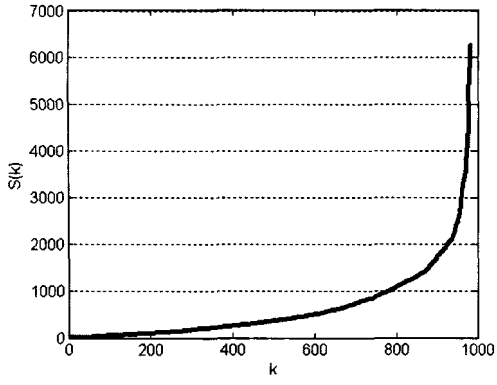

图 4a t(3) 分布的 Sum-plot 图

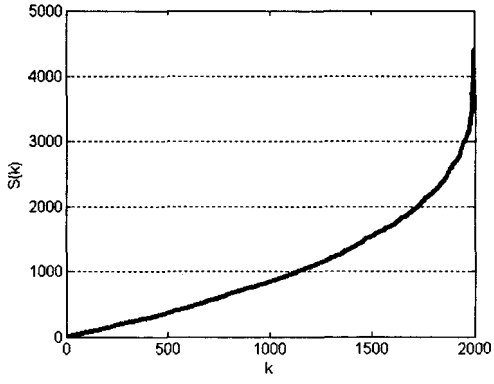

图 4b IGa(1.5,1) 分布的 Sum-plot 图

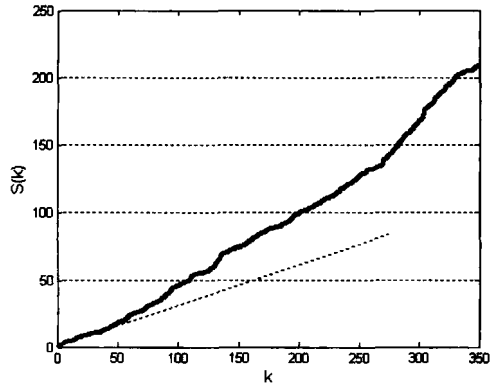

图 5a t(3) 分布 350 个较大顺序统计量的 Sum-plot 图

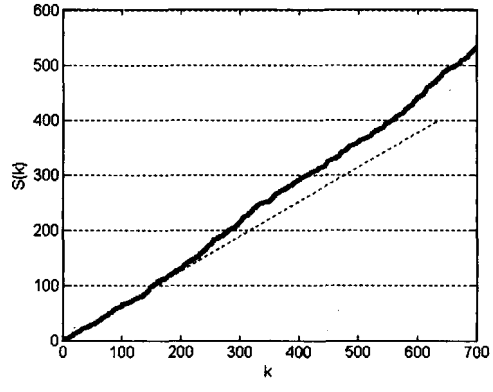

图 5b IGa(1.5,1)) 分布 700 个较大顺序统计量的 Sum-plot 图

由图 4 可以看出, Sum-plot 图仅在开始的区域内呈线性, 当数据数目很大时, 一次作图效果有时不是很清晰, 二次作图非常必要. 为了更准确地确定图形的线性部分, 我们再次作呈现线性部分的那些较大的顺序统计量的 Sum-plot 图. 考虑 t(3) 分布的前 350 个较大的顺序统计量和 IGa(1.5,1) 分布的前 700 个较大的顺序统计量, 作 Sum-plot 图. 由图 5 可以看出, 对

于 t(3) 分布, 当 $k=32$ 时, 图形偏离线性. 对于 IGa(1.5,1) 分布, 当 $k=160$ 时图形偏离线性. 利用 Hill 估计计算它们的尾部指数分别为: $H_{984,32}$=3.0024, $H_{2000,160}$=1.5174. 它们很接近真实值 3 和 1.5.

综上, Sum-plot 方法针对 Pareto 分布选取的 k 就是样本容量 n, 使用了样本数据的全部信息. 而对于 t(3) 分布和 IGa(1.5,1) 分布, 选取的 k 值分别为 32 和 160, 都分别在类似图 2 中.α 的置信区域对应的 k 值的范围内.

使用 Bootstrap 方法模拟时, 针对 Pareto(1.5) 和 t(3) 分布, 抽取样本容量为 1000 的样本, 选取 n_1 的范围从 500 到 850, 增量为 50, 随机抽取 Bootstrap 样本 2000 次. 针对 IGa(1.5,1) 分布, 抽取样本容量为 2000 的样本, 选取 n_1 的范围从 800 到 1800, 增量为 100, 随机抽取 Bootstrap 样本 1500 次.

使用 M-Bootstrap 方法模拟时, 同样针对 Pareto(1.5) 和 t(3) 分布, 抽取样本容量为 1000 的样本, 针对 IGa(1.5,1) 分布, 抽取样本容量为 2000 的样本, 随机抽取 Bootstrap 样本 20000 次, 参数按照 Hall 的建议选取 μ=2/3, β=1/2.

为便于比较, 我们将三种方法的模拟结果都列表如下 (见表 1):

表 1 三种方法用于 Pareto(1.5)、t(3)、IG(1.5) 分布的结果

估计方法	Stable(1.5) ($n=996$)		t(3) ($n=984$)		IGa(1.5,1) ($n=2000$)	
	α	k	α	k	α	k
sum-plot	1.5146	60	3.0024	32	1.5174	160
bootstrap	1.5699	248	3.1213	34	1.4140	359
M-bootstrap	1.5644	90	2.5581	69	1.4866	176

根据表 1 可以看出, 应用三种方法得到的结果是令人满意的. 相比之下, Sum-plot 方法的精确性优于两种 Bootstrap 方法. 从整体上看, 两种 Bootstrap 方法估计的结果误差也是比较小的, 都可以使用. 从 k 选择上看, 改进的 M-Bootstrap 方法更接近 Sum-plot 方法结果, 对重尾指数的估计在某些情况下优于 Bootstrap 方法, 特别是在计算量上明显优于 Bootstrap 方法. 所以, M-Bootstrap 方法是适用的, 有意义的. 两种 Bootstrap 方法个别情形下出现了较大的偏差, 这与方法本身的特点有关. 基于两个子样本的 Bootstrap 方法受异常值的影响, 我们所用的数据都是随机生成的, 不免有异常值的出现. Bootstrap 方法受样本容量的影响很大, 这也是出现偏差的原因.

5 沪深股市重尾性实证分析

本文选取 1991 年 5 月 6 日到 2006 年 9 月 29 日时间段上证综合指数 (共 3783 个数据) 和深证成份指数 (共 3802 个数据) 每日收盘价 p_t 作为基础数据, 股指收益 $R_t=\log(p_t/p_{t-1})$ 作为样本数据. 我们绘制了所选两个指数的股指收益图, 见图 6(a)、图 6(b). 图中的横轴表示以日为单位的时间, 为简便起见, 依序列号代替, 纵轴表示股指收益. 在 1992 年 5 月 21 日, 上海股市交易价格限制全部取消, 股市交易价格开始尝试由市场引导, 由于政策性影响上证指数首度跨越千点, 在全面放开股价的利好刺激下, 大盘直接跳空高开在 1260.32 点, 较前一天涨幅高达 104.27%. 为消除异常值得影响, 我们同时考虑去掉该值情形.

为了获得股指收益分布的主要特征, 我们首先计算样本的基本统计量: 均值, 标准差, 偏度和峰度 (见表 2). 计算结果表明, 上证综合指数股指收益和深证成份指数股指收益的偏度都为正, 峰度都明显大 3, 从偏度和峰度数值可以看出, 指数收益分布较正态分布有偏且具有明显的尖峰特征.

基本统计结果为股指收益偏离正态分布提供了证据, 但是为了得到更确切的结论, 需要对样本做正态性检验. 我们采用 SAS 软件提供的 Kolmogorov-Smirnov、Cramer-von Mises 和 Anderson-Darling 三种检验方法. 这三种检验方法的零假设都是正态分布, 如果检验的 p 值小于 0.05, 应当拒绝零假设, 认为该序列不服从正态分布, 否则接受零假设, 认为该序列服从正态分布. 检验的结果如表 3.

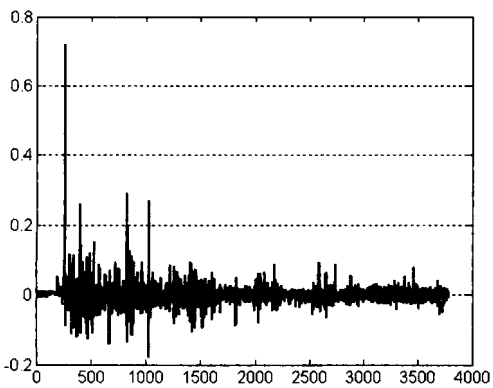

图 6a 上证综合指数收益率的波动图

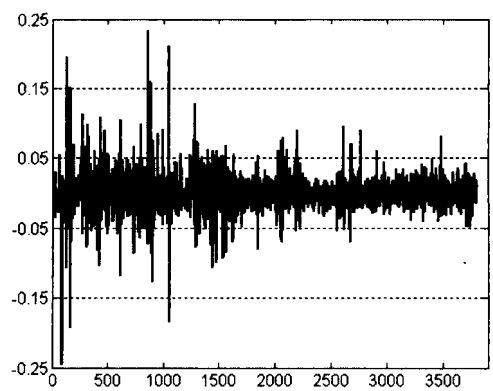

图 6b 深证成份指数收益率的波动图

表 2 指数股指收益的基本统计结果

指数	均值	标准差	偏度	峰度
上证综合指数	0.00073	0.026960	6.03199	149.29682
深证成份指数	0.00042	0.023110	0.54159	19.37672
上证综合指数 (去异)	0.00053	0.024298	1.42268	23.84946

表 3(a) 上证综合指数股指收益的正态性检验

检验	统计量		P 值	
Kolmogorov –Smirnov	D	0.158700	Pr> D	<0.0100
Cramer-von Mises	W-Sq	35.79628	Pr >W-Sq	<0.0050
Anderson-Darlin	A-Sq	196.12220	Pr >A-Sq	<0.0050

表 3(b) 深证成分指数股指收益的正态性检验

检验	统计量		P 值	
Kolmogorov –Smirnov	D	0.10783	Pr > D	< 0.0100
Cramer-von Mises	W-Sq	17.79667	Pr >W-Sq	<0.0050
Anderson-Darlin	A-Sq	100.3423	Pr >A-Sq	<0.0050

表 3(c) 证综合指数股指收益 (去异) 的正态性检验

检验	统计量		P 值	
Kolmogorov –Smirnov	D	0.13962	Pr > D	< 0.0100
Cramer-von Mises	W-Sq	26.61371	Pr >W-Sq	<0.0050
Anderson-Darlin	A-Sq	147.93200	Pr >A-Sq	<0.0050

从表 3 可以看到, 对于上证综合指数股指收益 (包括去掉异常值情形) 和深证成分指数股指收益两组样本, 三种检验方法的 p 值都小于 0.01, 说明股指收益的正态分布假设不能成立.

根据指数股指收益的基本统计结果知股指收益呈现出尖峰特征, 为了更直观地理解其内涵, 我们可以绘制所选两个指数股指收益 (包括去掉异常值的上证综合指数股指收益) 的概率密度直方图和分位数 - 分位数图 (QQ-plot). 在 QQ-plot 图中我们选取正态分布作为参考分布, 即以标准正态分布的分位数作横轴, 以数据的分位数作纵轴. 根据直方图和 QQ-plot 可以明显看出两个指数股指收益 (包括去掉异常值的上证综合指数股指收益) 分布都比正态分布的尾部厚, 进一步表明服从重尾分布 (参见图 7– 图 8).

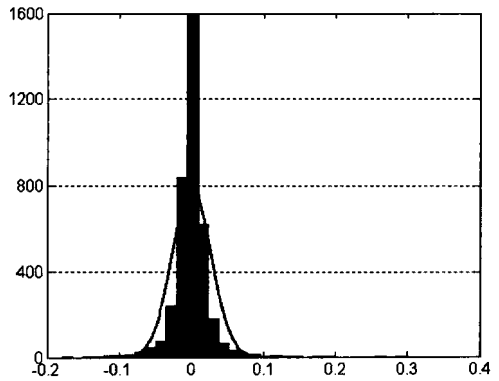

图 7 上证综合指数收益率的概率密度直方图

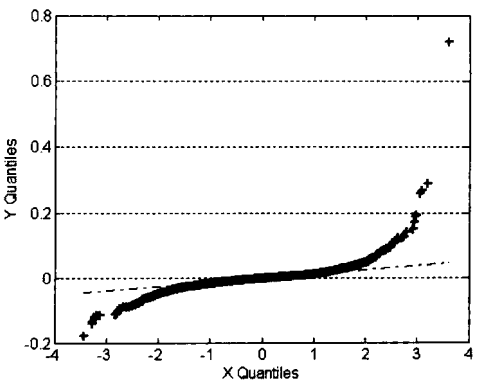

图 8 上证综合指数收益率的 QQ-plot 图

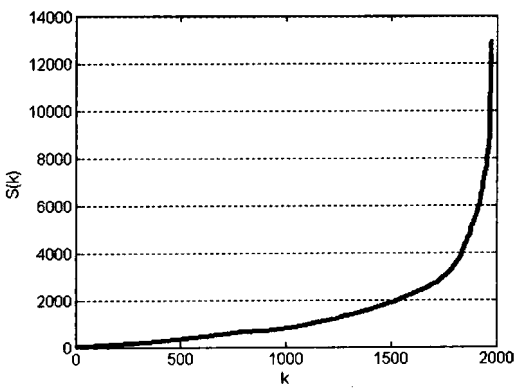

图 9 上证综合指数右尾的 Sum-plot 图

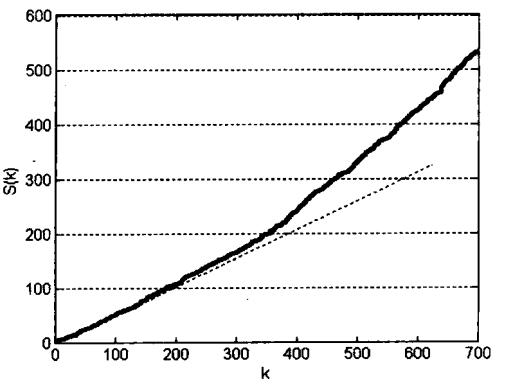

图 10 右尾 700 个较大顺序统计量的 Sum-plot 图

首先我们应用 Sum-plot 和 Hill 估计相结合来计算股指收益序列的重尾指数, 从而量化

重尾程度. 由于两个指数股指收益分布都具有不同程度的偏态性, 所以我们分别讨论它们的左尾 (高损失) 和右尾 (高回报) 特征.

根据样本作 Sum-plot 图, 从 Sum-plot 图中呈线性部分, 初步选择一定数量的较大顺序统计量再进行二次作图, 根据新作 Sum-plot 图我们确定上证综指股指收益右尾 1976 个数据中 142 个数据位于尾部, 利用 Hill 估计计算得到尾部指数 α_2=1.9699. 从确定证综指股指收益左尾 1807 个数据中 88 个数据位于尾部, 利用 Hill 估计计算得到尾部指数 α_1=2.8828. 由于重尾指数 α 越小说明尾部越厚, 所以计算得到的结果与检验结果一致, 对于上证综指股指收益都是右尾厚于左尾. 为了直观理解确定过程, 我们以上证综指股指收益右尾的 Sum-plot 图为例 (见图 9- 图 10).

表 4 三种方法分别应用于上证综指和深证成份指数股指收益率的结果

方法	上证综指(左尾) (n=1807)		上证综指(右尾) (n=1976)		上证综指(右尾去异) (n=1975)		深证成指(左尾) (n=1952)		深证成指(右尾) (n=1849)	
	α	k	α	k	α	k	α	k	α	k
sum plot	2.8828	88	1.9699	142	2.0388	141	3.3154	33	2.5643	86
bootstrap	3.0220	69	2.1005	64	2.5526	6	3.1885	35	2.5700	52
M-bootstrap	2.5328	122	1.9707	113	2.0650	113	2.9147	75	2.5636	73

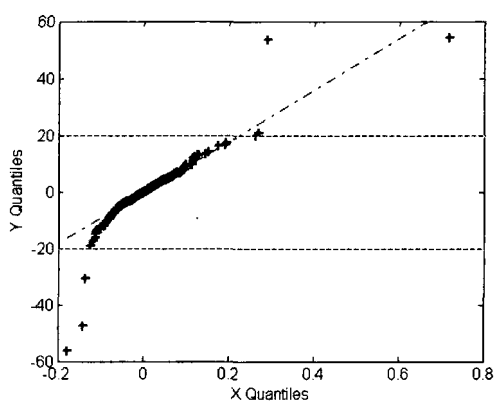

图 11a 基于 t(2) 分布的上证指数 QQ-plot 图

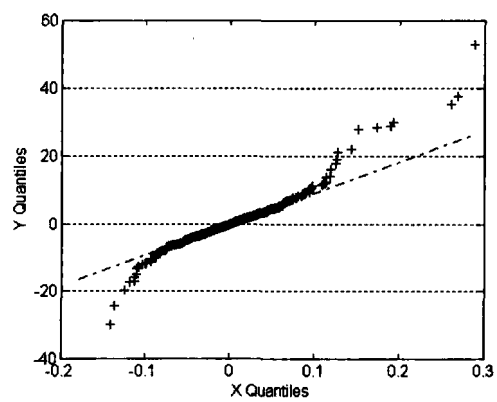

图 11b 基于 t(2) 分布的上证指数 (去异)QQ-plot 图

上证综指股指收益波动图 (图 6) 显示上证综合指数出现了极端值 (异常值), 时间是 1992 年 5 月 21 日, 我们已将除去异常值后的新样本的基本统计量结果列在表 2 中. 由表 2 可以看出异常值对偏度, 峰度的影响很大. 除去这个异常值, 样本的峰度从 149 下降到 23, 偏度也降低了. 由表 3(c) 的正态性检验结果和概率密度直方图及 QQ-plot 图得出除去异常值后, 上证综合指数股指收益还是重尾的. 因为异常值是正的, 所以我们只需要单独考虑除去异常值后的上证综合指数股指收益的右尾.

通过作 Sum-plot 图, 我们确定除去异常值后上证综合指数右尾 1975 个数据中 141 个数据位于尾部, 利用 Hill 估计计算得到尾部指数 α_2=2.0388. 与不去掉异常值情形相比差别不大, 可以认为 Sum-plot 方法不受异常值得影响, 具有稳健性.

我们利用同样的方法估计深证成份指数股指收益的重尾指数.

通过作 Sum-plot 图我们确定深证成份指数股指收益左尾的 1952 个数据中 33 个数据位于尾部, 利用 Hill 估计计算得到尾部指数 α_1=3.3154. 确定右尾 1849 个数据中 81 个数据位于尾部, 利用 Hill 估计计算得到尾部指数 α_2=2.4261. 由于尾部指数 α 越小说明尾部越厚, 所以计算得到的结果与检验结果一致, 都是右尾厚于左尾.

我们再分别使用 Bootstrap 方法和 M-Bootstrap 方法, 结合 Hill 估计来对估计沪深股市股指收益的重尾指数. 使用 Bootstrap 方法时, 我们选取 n_1 的范围从 800 到 1800, 增量为 100, 随机抽取 Bootstrap 样本 1500 次. 利用 M-Bootstrap 方法时, 随机抽取 Bootstrap 样本 20000 次. 同时选择参数 μ=2/3, β=1/2. 计算结果列于表 4.

由于上证综合指数股指收益右尾的尾部指数接近 2, 应当可以用比正态分布具有更重尾的 t(2) 分布来拟合, 我们可以通过 QQ- 图证明这一点. 由图 11 可以看出, 当我们用 t(2) 分布来对上证综合指数股指收益拟合时, 除去异常值后的结果更好一些. 实际上, 我们并不希望除去异常值, 正是由于极端值的出现, 才使重尾的特征更明显. 结论表明研究上证综合指数股指收益时使用 t(2) 分布即可.

综上实证分析, 对我国上证综指和深证成指的股指收益序列的分布性质给出了基本判断, 比较精确地估出了分布的重尾指数, 为进一步研究我国股市的股指收益、金融市场风险、金融资产定价以及金融创新提供了依据. 同时实证分析也为三种重尾指数估计方法提供了比较, 我们得到以下一些结论:

1) 总体上看, 应用 Sum-plot 方法和 Bootstrap 方法都能够得到令人满意的结果, Sum-plot 方法优于 Bootstrap 方法和 M-Bootstrap 方法;

2) 应用 M-Bootstrap 计算量比 Bootstrap 方法要小得多, 而且效果比 Bootstrap 方法要好;

3) 当出现大的极端值 (异常值) 时, Sum-plot 方法和 M-Bootstrap 方法选择 k 几乎不受影响, 而利用两个子样的 Bootstrap 方法的结果受异常值的影响较大.

所以, 在估计股指收益序列的重尾指数时, 同时使用 Sum-plot 方法和 M-Bootstrap 方法, 既能得到比较精确的结果, 又能相互验证.

6 结论

本文研究了估计重尾指数时选取 k 的两类方法: Sum-plot 方法和 Bootstrap 方法, 提出了一个对 Hall 方法的有意义的改进 M-Bootstrap 方法. 通过模拟以及实证分析, 发现利用 M-Bootstrap 方法选取的 k 应用到 Hill 估计中所计算的重尾指数有很高的精确性, 而且这种方法方便省时. Sum-plot 方法在很大程度上克服了 Hill-plot 方法 qq-plot 方法选取 k 所遇到的困难, 并且它所选取的 k 运用到重尾指数的计算上, 误差很小. 基于两个子样本的 Bootstrap 方法也有较好的性质, 估计精确性比较高, 但是受异常值的影响较大.

利用这些方法对沪深股市的实证分析, 证明了两市的股指收益率具有重尾性. 计算出的重尾指数表明两市的股指收益率都是右尾厚于左尾, 具有右偏态性, 这与检验的结果是一致的. 而且我们还发现它们的右尾指数都小于 3, 这反映出我国的股票市场整体上是高回报的.

重尾分布的一些特征通过重尾指数反映出来, 金融风险的度量也与重尾指数有关, 因而,

重尾指数的估计显的尤为重要. 而重尾指数的估计又与所用较大顺序统计量的个数密切相关. 到目前为止, 重尾分布的重尾指数估计方法都局限于随机变量独立的情况下, 而实际问题中的一些时间序列常表现出相关性, 比如在不是有效的金融市场中的金融时间序列具有长记忆性. 有必要推广到非独立情况, 这将是我们以后努力的方向. 今后我们应该从图形表现出的一些特征 (比如凸凹性) 来判断其底分布所属于的类型, 这样会更有助于我们选取更优的 k, 能够更精确地估计重尾指数. 这对研究在重尾分布下的风险管理有着重要的意义.

参考文献

[1] Zipf G. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology[M]. MA: Addison-Wesley Press, 1949.

[2] Fama E and R Roll. Some properties of symmetric stable distributions[J]. Journal of the American Statistical Association, 1968, 63: 817-836.

[3] Press S. Estimation in univariate and multivariate stable distributions[J]. Journal of the American Statistical Association, 1972, 67(340): 842-846.

[4] Zolotarev V. One-dimensional stable distributions[J]. Translations of Mathematical monographs (American Mathematical Society), 1986, 65.

[5] Pickands. J Statistical inference using extreme order statistics[J]. Annals of Statistics, 1975, 3: 119-131.

[6] Hill B. A simple general approach to inference about the tail of a distribution[J]. Annals of Statistics, 1975, 3: 1163-1174.

[7] Csörgö S, P Deheuvels and D Mason. Kernel estimates of the tail index of a distribution[J]. Annals of Statistics, 1985, 13(3): 1050-1077.

[8] Smith R. Estimating tails of probability distributions[J]. Annals of Statistics, 1987, 15: 1174-1207.

[9] Smith R. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone[J]. Statistical Science, 1989, 4: 367-393.

[10] Dekkers A and L de Haan. On the estimation of the extreme value index and large quantile estimation[J]. Annals of Statistics, 1989, 17(4): 1795-1832.

[11] Kratz M and Resnick S. The qq-estimator and heavy tails[J]. Comm Statist Stochastic Models, 1996, 12: 699-724.

[12] Beirlant J, Vynckier P and Teugels J. Tail index estimation, Pareto quantile plots, and regression diagnostics[J]. Journal of the American Statistical Association, 1996, 91: 1659-1667.

[13] Beirlant J, Vynckier P and Teugels J. Excess functions and estimation of the extreme value index[J]. Bernoulli, 1996, 2: 293-318.

[14] Feuerverger A and Hall P. Estimating a tail exponent by modelling departure from a Pareto distribution[J]. Annals of Statistics, 1999, 27: 760-781.

[15] Resnick S and C Stăică Smoothing the hill estimator[J]. Advances in Applied Probability, 1997. 29: 271-293.

[16] Mandelbrot B. The variation of certain speculative prices[J]. Journal of Business, 1963, 36: 304-419.

[17] Sousa B. A contribution to the estimation of the tail index of heavy-tailed distributions. Ph.D. thesis (in The University of Michigan), www.utstat.toronto.edu/~desousa, 2002.

[18] B Efron. Bootstrap methods: Another look at the Jackknife[J]. Annals of Statistics, 1979, 7(1): 1-26.

[19] Hall P. Using the bootstrap to estimate mean square error and select smoothing parameters in non-parametric problems[J]. Journal of Multivariate Analysis, 1990, 32: 177-203.

[20] Jef Caers and Jozef Van Dyck. Nonparametric tail estimation using a double bootstrap method[J]. Computational Statistics & Data Analysis, 1999, 29(2): 191-211.

[21] Danielsson J, L de Haan, L Peng et al. Using a bootstrap method choose the sample fraction in tail index estimation[J]. Journal of Multivariate Analysis, 2001, 76:226-248.

[22] 刘维奇, 赫莫迪, 邢红卫. 选择重尾阀值 k 的 Bootstrap 方法 [J]. 山西大学学报 (自然科学版), 2010,4.

Tail Index Estimation of Heavy-tailed Distribution and Empirical Analysis of China's Stock Markets

LIU Wei-qi[1,2], HE Ying-di[2,3], CHEN Lin[2,4]

(1. Institute of Management Science and Engineering, Shanxi University, Taiyuan 030006, China)
(2. School of Mathematics Science, Shanxi University, Shanxi 030006, China)
(3. Maoming Vocational Technical College, Maoming 525000, China)
(4. Engineering College of Shanxi University, Taiyuan 030013, China)

Abstract: Estimating the tail index α of a heavy-tailed distribution depends on the choice of the number k of upper order statistics used in the estimation. In this paper, we introduce two different kinds of methods which are used to choose the value k in the estimating α, Sum-plot method and Bootstrap method. And also improve the Bootstrap method proposed by Hall, it is called M-Bootstrap method. Firstly, using the above methods, we simulate known distributions and prove feasibility of these methods by Monte-Carlo method. Then, we make empirical analysis basing on Shanghai and Shenzhen's stock index data. These results show that China's stock index returns rate is of thick tail, and expose right skewness, that is, right tail is heavier on left tail. By comparing results obtained in several different methods, indicate that Sum-plot method and M-Bootstrap method is more precise in estimating heavy-tailed index and is immune to anomalous value.

Keywords: heavy-tailed distribution; tail index of heavy-tailed distribution; Hill estimation; sum-plot method; bootstrap method