山西大学学报(自然科学版)33(4): 508~ 512, 2010
Journal of Shanxi University(Nat Sci Ed)

文章编号: 0253-2395(2010)04-0508-05

选择重尾阈值 k 的 Bootstrap方法

刘维奇[1,2], 赫英迪[2,3], 邢红卫[2]
(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)

摘 要: 详细讨论了重尾指数估计中选取 k 的 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. 并利用上述三种方法对已知重尾分布进行 Monte-Carlo模拟, 研究它们的可行性, 比较它们的稳健性, 改进的 M-Bootstrap方法对重尾指数的估计在某些情况下优于 Bootstrap方法.

关键词: 重尾指数; 重尾阈值; Sum-plot方法; Bootstrap方法; M-Bootstrap方法
中图分类号: O212 文献标识码: A

重尾指数估计方法总体上分为参数估计和半参数估计, 都与重尾阈值或估计中所用次序统计量的个数 k 有关. k 的选取关系到估计的精确性, k 的偏大或偏小都会造成估计的极大误差.

学者们从理论上提出了许多选取 k 的方法. 其中一类是作图法, 比如 Hill[1]提出的 Hill-plot, Kratz和 Resnick[2]提出的 qq-plot, Beirlant等[3]提出的 Pareto分位数图, Resnick和 Starica[4]给出的对 Hill-plot改进的 smoo-Hill-plot以及 de-Haan和 Resnick[5]给出的对 Hill-plot改进的 Alt-Hill-plot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所有情况的重尾分布. 像 Hill-plot, qq-plot, 当随机变量服从 Pareto分布时, 这两种方法表现出十分优良的性质, 能够很容易选取 k 值. 一旦随机变量不服从 Pareto分布, 而是广义 Pareto分布时, 它们却不能很好地选取 k, 甚至无法选取 k. Pareto分位数图, smoo-Hill-plot和 Alt-Hill-plot相对于 Hill-plot估计精度稍高一些, 但是也不能对所有的重尾分布较好地选择 k. Sousa[6]在其博士论文中提出的 Sum-plot方法在一定程度上克服了前几种方法中选取 k 所遇到的困难, 而且具有比较好的性质. 但是由于 Sum-plot方法是以观察图形得到 k, 因此选择 k 有一定的猜测性, 因而会对重尾指数估计造成一定误差. 另一类方法就是以估计重尾指数的均方误差(MSE)最小为标准来确定 k, 最优的 k 应该与均方误差一致. 理论上 MSE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. 只有权衡方差和偏差使 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, 不能直接应用到实际问题中. 基于此, 1990年 Hall[7]提出了利用 Bootstrap方法来选取 k, Danielsson[8]在 2001年又对 Hall的方法作了进一步改进, Gomes和 Oliveira[9]在 2001年给出了一个选取 Bootstrap方法子样本的准则, Gomes等①在 2009年给出了针对降偏差重尾指数估计的 Bootstrap方法. 由于该方法计算量很大, 有必要在保证估计特性的前提下提高估计的收敛速率以减少计算量.

1 Sum-plot方法

Sum-plot方法[6]是基于 $\{(k, S_k),\ 1\leqslant k\leqslant n\}$ 应该是一条直线的理论依据来选取 k. Sousa通过对不同样本容量的不同分布进行模拟, 得出无论是分布的尾部指数 $0<\alpha<2$ 还是 $\alpha\geqslant 2$, Sum-plot方法对绝大多数分布

收稿日期: 2010-07-16; 修回日期: 2010-07-30
基金项目: 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)
作者简介: 刘维奇(1963-), 男, 山西忻县人, 教授, 博士生导师, 主要从事金融工程和时间序列等领域的研究, E-mail: liuwq@sxu.edu.cn

① Gomes M I, Mendonca S, Pestana D. The bootstrap methodology and adaptive reduced-bias tail index and Value-at-Risk estimation. Working paper, 2009.

而言都较其它方法优越, 并且不受样本异常值影响, 即具有稳健性. 这里随机变量

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

其中 $X_n^{(1)} \geqslant X_n^{(2)} \geqslant \cdots X_n^{(k+1)}$ 为次序统计量.

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

$$S_i = \beta_0 + \beta_1 i + \varepsilon_i,\ i = 1, 2, \cdots, k. \tag{2}$$

容易发现参数 α^{-1}的估计值等于回归模型的斜率 β_1, 即

$$\hat{\alpha}_{n,k}^{-1} = \beta_1 = \frac{k}{k-1} H_{n,k}^{-1} - \frac{1}{k-1} \log X_n^{(1)} \tag{3}$$

进一步, 如果 $\beta_0 = 0$, 则 $\hat{\alpha}_{n,k}^{-1} = \beta_{LS} = H_{n,k}^{-1}$, 其中 $H_{n,k}^{-1}$就是 Hill估计.

由于 Sum-plot方法需要观察以坐标 $\{(k, S_k),\ 1 \leqslant k \leqslant n\}$画成的散点图在哪一点偏离直线, 因此选择的 k 有一定的猜测性, 因而会对重尾指数估计造成不可避免的误差.

2 *Danielsson-Bootstrap* 方法

Danielsson等[8]对 Hall的方法作了改进, 使用新的统计量 $M_n(k)$来代替 $\gamma_n(k)$. 引入统计量

$$M_n(k) = \frac{1}{k} \sum_{i=1}^{k} (\log X_n^{(i)} - \log X_n^{(k+1)})^2,\ 1 \leqslant k \leqslant n. \tag{4}$$

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 AMSE和极小化 $AsyE(M_n(k) - 2(\gamma_n(k))^2)^2$可以得到同阶量的 k(相对于 n). 因此, 根据 Bootstrap子样本 $X_{n_1}^*$, 选用统计量:

$$Q(n_1, k_1) = E((M_{n_1}^*(k_1) - 2(\gamma_{n_1}^*(k_1))^2)^2 \mid X_n), \tag{5}$$

其中 $M_{n_1}^*(k_1) = \frac{1}{k_1} \sum_{i=1}^{k_1} (\log X_n^{(i)*} - \log X_n^{(k+1)*})^2$. 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 $X_{n_2}^*$, $n_2 = n_1^2/n$, 然后利用与确定 k_1 相同的程序来确定 k_2, 再利用 k, k_1 和 k_2 之间的关系

$$k = \frac{k_1^2}{k_2} \left(\frac{(\log k_1)^2}{(2\log n_1 - \log k_1)^2} \right)^{\frac{\log n_1 - \log k_1}{\log n_1}} \tag{6}$$

来确定 k.

3 *M-Bootstrap* 方法

我们受 Danielsson等[8]提出的 Bootstrap方法的启发, 用 γ的相合估计 $\tilde{\gamma}_n(k)$代替 $\gamma_n(k)$, 渐近均方误差变为

$$AMSE_M(n_1, k_1) = E((\gamma_{n_1}^*(k_1) - \tilde{\gamma}_n(k))^2 \mid X_n),\ 1 \leqslant k < n \tag{7}$$

根据 Bootstrap子样本 $X_{n_1}^*$, 通过极小化 $AMSE_M(n_1, k_1)$和关系 $k = k_1 (n/n_1)^{\mu}$来确定 k_1 与 k.

定理 1 假设 $k \to \infty$, $k/n \to 0$ $k(n)$由 $AMSE(n, k)$最小确定. 则

$$k = \frac{n}{S^{-1}(\gamma^2(1-\rho)^2/n)} (1 + o(1)),\ n \to \infty, \tag{8}$$

S^{-1}是函数 S的反函数, $A^2(t) = \int_t^{\infty} S(u)\,du\,(1 + o(1))$, $t \to \infty$.

假设 $A(t) = ct^{\rho}$, $c \neq 0$, $\rho > 0$, 则

$$k = H(\rho) n^{\mu} (1 + o(1)),\ \mu = 2\rho/(2\rho - 1) \tag{9}$$

定理 2 假设 $k_1 \to \infty$, $k_1/n_1 \to \infty$. 假设 $A(t) = ct^{\rho}$, $c \neq 0$, $\rho < 0$, $n_1 = O(n^{1-\varepsilon})$ $(0 < \varepsilon < 1)$, 由 $AMSE_M(n_1, k_1)$最小确定 k_1. 则

$$k_1 = H(\rho) n_1^{\mu} (1 + o(1)),\ \mu = 2\rho/(2\rho - 1) \tag{10}$$

由定理 1和定理 2可知, k 与 n, k_1 与 n_1 存在同样的幂指数关系式. 这与 Hall所预设的关系一致. 所以我们仍旧取 $\mu=\frac{2}{3}$, $\beta=\frac{1}{2}$, 由 $k=k_1(\frac{n}{n_1})^{\mu}$ 来确定 k. 我们取 $\mu=\frac{2}{3}$, 无形中假设了二阶形状参数 $\rho=-1$, 这证实了 Hall的 Bootstrap方法与 $\rho=-1$ 有关.

随机变量 $Y_1, Y_2, \cdots, Y_n$ 是 iid, 其共同分布为 $G(y)=1-y^{-1}(y\geqslant 1)$, $Y_{n,1}\geqslant\cdots\geqslant Y_{n,n}$ 是 $Y_1, Y_2, \cdots, Y_n$ 的顺序统计量. $\{X_{n,1}\}_i^n \overset{d}{=} \{U(Y_{n,i})\}_{i=1}^n$, 其中 $U(t)=(\frac{1}{1-F})^{-}(t)$.

引理 1 $0<k<n$, 且 $k\to\infty$, 则有

(1) $n\to\infty$, $\frac{Y_{n,k}}{(n/k)}\xrightarrow{p}1$

(2) $n\to\infty$, (P_n, Q_n) 渐近正态, 它们的均值为 0, 方差分别为 1, 2, 协方差为 4, 其中

$$P_n=\sqrt{k}\{\frac{1}{k}\sum_{i=1}^{k}\log Y_{n,i}-\log Y_{n,k+1}-1\},\ Q_n=\sqrt{k}\{\frac{1}{k}\sum_{i=1}^{k}(\log Y_{n,i}-\log Y_{n,k+1})^2-2\}.$$

定理 1的证明: $U(t)$ 的定义等价于正则变化函数 $|\log U(t)-\gamma\log t-C_0|$ 以指数 ρ 正则变化, 其中 C_0 为常数. 令 $A(t)=\rho(\log U(t)-\gamma\log t-C_0)$. 由 Potter不等式, 可得对任意 $0<\varepsilon<1$, 存在 $t_0>0$, 对于 $t_0>0$, $tx\geqslant t_0$ 有,

$$(1-\varepsilon)x^{\rho}e^{-\varepsilon|\log x|}-1\leqslant\frac{\log U(tx)-\log U(t)-\gamma\log X}{A(t)/\rho}\leqslant(1+\varepsilon)x^{\rho}e^{\varepsilon|\log X|}-1. \tag{11}$$

用 $Y_{n,k}$ 代替 t, $Y_{n,i}/Y_{n,k+1}$ 代替 x 迭代不等式 $(i=1, 2, \cdots, k)$, 然后乘以 $\frac{1}{k}$, 得到

$$\gamma\approx\gamma+\frac{\gamma P_n}{\sqrt{k}}+\rho^{-1}A(Y_{n,k+1})(1\pm\varepsilon)\{\frac{1}{k}\sum_{i=1}^{k}(\frac{Y_{n,i}}{Y_{n,k+1}})^{\rho\pm\varepsilon}-1\}.$$

又

$$\sum_{i=1}^{k}\frac{Y_{n,i}}{Y_{n,k+1}}\overset{d}{=}\sum_{i=1}^{k}Y_i,$$

而 $Y_1, \cdots, Y_k$ iid具有共同分布函数 $1-\frac{1}{y}$, 于是由弱大数定律得

$$\gamma_n\approx\gamma+\frac{\gamma P_n}{\sqrt{k}}+\rho^{-1}(1\pm\varepsilon)(\frac{1}{1-\rho\mp\varepsilon}-1)A(Y_{n,k}),$$

即

$$\gamma_n\approx\gamma+\frac{\gamma P_n}{\sqrt{k}}+(1-\rho)^{-1}A(\frac{n}{k})+O_p(A(\frac{n}{k})),$$

$$AsyE(\gamma_n-\gamma)^2\approx\frac{\gamma^2}{k}+\frac{A^2(n/k)}{(1-\rho)^2} \tag{12}$$

我们求 (12)中右边的最小值点, 得到定理 1的结论, 定理证毕.

定理 2的证明: 令 G_n 表示独立变量的均匀分布的经验分布函数. 令 n 足够大, $n_1=O(n^{1-\varepsilon})$, 则有

$$Y_2\leqslant\sup_{0<t\leqslant n_1(\log n_1)^2}tG_n^{-}(\frac{1}{t})\leqslant 2,\ a\ s \tag{13}$$

$$\sup_{t\geqslant 2}|\sqrt{t}(G_n(\frac{1}{t})-\frac{1}{t})|\leqslant\frac{\log n}{\sqrt{n}},\ a\ s$$

于是

$$\sup_{4\leqslant t\leqslant n_1(\log n_1)^2}\left|\frac{\sqrt{1}}{\sqrt{G_n^{-}(\frac{1}{t})}}[G_n(G_n^{-}(\frac{1}{t}))-G_n^{-}(\frac{1}{t})]\right|\leqslant\frac{\log n}{\sqrt{n}}$$

因此, 对所有的 $4\leqslant t\leqslant n_1(\log n_n)^2$,

$$|tG_n^{-}(\frac{1}{t})-1|\leqslant\frac{2\sqrt{t}\log n}{\sqrt{n}},\ a\ s \tag{14}$$

用 F_n 表示 X_n 的经验分布函数. $v(t)=G_n^-\left(1-\frac{1}{t}\right)$, 由 (11), (13), (14)得,

$$|\log y|\leqslant 2|y-1|,\ \frac{1}{2}\leqslant y\leqslant 2,$$

$$|y^{-e}-1|\leqslant(-\rho)(2^{\rho-1}\vee 2^{1+\rho})|y-1|,\ \frac{1}{2}\leqslant y\leqslant 2,$$

$$\log U_n(t)=\log F_n^-\left(1-\frac{1}{t}\right)\overset{d}{=}$$

$$\log F_n^-\left(G_n^-\left(1-\frac{1}{t}\right)\right)=\log v\left(\frac{1}{1-G_n^-\left(1-\frac{1}{t}\right)}\right)\overset{d}{=}\log v\left(\frac{t}{tG_n^-\left(\frac{1}{t}\right)}\right).$$

所以对任意的 $0<\varepsilon<1$, 总存在 $t_0>4$, 对于 $t_0<t<n_1(\log n_1)^2$ 及 $t_0<tx<n_1(\log n_1)^2$, 有

$$\frac{\log U_n(tx)-\log U_n(t)-\gamma\log x}{A(t)/\rho}\overset{d}{=}$$

$$\log\frac{\frac{\log U\left(\frac{tx}{txG_n^-\left(\frac{1}{tx}\right)}\right)-\log U(tx)-\gamma\log\left(\frac{1}{txG_n^-\left(\frac{1}{tx}\right)}\right)}{A(tx)/\rho}\frac{A(tx)}{A(t)}+\frac{\gamma\log\left(\frac{1}{txG_n^-\left(\frac{1}{tx}\right)}\right)}{A(t)/\rho}-\frac{\gamma\log\left(\frac{1}{tG_n^-\left(\frac{1}{t}\right)}\right)}{A(t)/\rho}-$$

$$\frac{\log U\left(\frac{t}{tG_n^-\left(\frac{1}{t}\right)}\right)-\log U(t)-\gamma\log\left(\frac{1}{tG_n^-\left(\frac{1}{t}\right)}\right)}{A(t)/\rho}+\frac{\log U(tx)-\log U(t)-\gamma\log x}{A(t)/\rho}\leq$$

$$\left[(-\rho)(2^{-\rho+1}\vee 2^{\rho+3})+2\left|\frac{\gamma\rho}{A(t)}\right|\right]\frac{2\sqrt{t}\log n}{\sqrt{n}}(\sqrt{x}+1)+$$

$$(1+9\varepsilon)(1+\varepsilon)x^{\rho}\exp(\varepsilon|\log x|)-1+7\varepsilon \tag{15}$$

同理

$$\frac{\log U_n(tx)-\log U_n(t)-\gamma\log x}{A(x)/\rho}\geqslant-\left[(1-\rho)(2^{-\rho+1}\vee 2^{\rho+3})+2\left|\frac{\gamma\rho}{A(t)}\right|\right]\frac{2\sqrt{t}\log n}{\sqrt{n}}(\sqrt{x}+1)+$$

$$(1-9\varepsilon)(1-\varepsilon)x^{\rho}\exp(-\varepsilon|\log x|)-1-7\varepsilon \tag{16}$$

用 Y_{n_1,k_1+1}, $Y_{n_1,i}$ $(i=1,\cdots,k_1)$分别代替 t和 tx, 则不等式 (15), (16)是以概率成立的. 于是有

$$4\leqslant Y_{n_1,i}\leqslant Y_{n_1,n_1}\ (i=1,\cdots,k_1)$$

以概率成立.

$$\frac{Y_{n_1,n_1}}{n_1(\log n_1)^2}\xrightarrow{p}0,\ n_1\to\infty,\ k_1/n_1\to 0$$

我们极小化 $E((\overset{*}{\gamma}_{n_1}(k_1)-\gamma_n(k))^2|X_n)$.

由定理 1的证明过程可以得到

$$\overset{*}{\gamma}_{n_1}(k_1)\overset{d}{=}\gamma+\frac{\gamma P_{n_1}}{\sqrt{k_1}}+d_1A(Y_{n_1,k_1+1})+o_p(A(n_1/k_1))+O\left(\frac{\log n\sqrt{n_1/k_1}}{\sqrt{n}}\right),$$

又 $\overline{\gamma}_n$ 是 γ的相合估计, $\overline{\gamma}_n=\gamma+o_p(A(n_1/k_1))$, 又 $\log n\sqrt{n_1/k_1}/\sqrt{n}=o(1/\sqrt{k_1})$. 定理 2得证, 定理证毕.

4 Monte-Carlo模拟

为了更好地说明问题, 我们选用三种熟知的重尾分布, 稳定分布 Stable(1.5)分布、t分布 t(3)以及逆 Γ 分布 IGa(1.5, 1), 分别采用 Sum-plot方法、Danielsson等提出的 Bootstrap方法 (D-Bootstrap方法)和改进的 Bootstrap方法 (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. 为便于比较, 我们将三种方法的模拟结果列表如下 (P512见表 1).

表 1 三种方法用于 t, Cauchy, Fréchet, 逆 Ga, Burr和 Pareto的结果

Table 1 Results by the three methods on t, Cauchy, Fréchet, Inverse-Gamma, Burr and Pareto distributions

分布	Stable(1.5)		t(3)		逆 Ga(1.5, 1)	
方法	k	$\hat{\alpha}$	k	$\hat{\alpha}$	k	$\hat{\alpha}$
Sum-plot方法	60	1.514 6	32	3.002 4	75	1.489 9
D-Bootstrap方法	248	1.569 9	34	3.121 3	367	1.388 0
M-Bootstrap方法	90	1.564 4	69	2.558 1	176	1.488 6

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.

参考文献:

[1] HILL B. A Simple General Approach to Inferece about The Tail of a Distribution[J]. *Annals of Statistics*, 1975, **3**: 1163-1174.

[2] KRATZ M, RESNICK S. The qq-estimator and Heavy Tails[J]. *Stochastic models*, 1996, **12**(4): 699-724.

[3] BEIRLANT J, VYNCKIER P, TEUGELS J L. Tail Index Estimation, Pareto Quantileplots, and Regression Diagnostics[J]. *Journal of the American Statistical Association*, 1996, **436**: 1659-1667.

[4] RESNICK S, STARICA C. Smoothing the Hill Estimator[J]. *Advances in Applied Probability*, 1997, **29**: 271-293.

[5] DREES H, HAAN L D, RESNICK S. How to Make a Hill Plot[J]. *Annals of Statistics*, 2000, **28**: 254-274.

[6] SOUSA B. A Contribution to the Estimation of the Tail Index of Heavy-tailed Distributions[D]. The University of Michigan, 2002.

[7] 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.

[8] DANIELSSON J. Using a Bootstrap Method Choose the Sample Fraction in Tail Index Estimation[J]. *Journal of Multivariate Analysis*, 2001, **76**: 226-248.

[9] GOMES M I, OLIVEIRA O. The Bootstrap Methodology in Statistics of Extremes-choice of the Optimal Sample Fraction[J]. *Extremes*, 2001, **4**(4): 331-358.

Bootstrap Method in Selecting Heavy-tailed Threshold k

LIU Wei-qi[1,2], HE Ying-di[2,3], XING Hong-wei[2]

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

Abstract: We discuss the Sum-plot method and Bootstrap method in selecting k in heavy-tailed index estimation, and improve the Bootstrap method proposed by Hall, known as the M-Bootstrap Method. The three methods were used to study the known heavy-tailed distributions by Monte-Carlo simulation technology, including their feasibility. Moreover, their robustness was compared. The M-Bootstrap method was better than the Bootstrap method in some cases for heavy-tailed index estimation.

Key words: heavy-tailed index; heavy-tailed threshold; Sum-plot method; Bootstrap method; M-Bootstrap method