有界范围内的幂律：从样本估计下界和上界

Data

Huan-Xiang Zhou

伊利诺伊大学芝加哥分校化学系和物理系，美国伊利诺伊州芝加哥市，邮编 60607

电子邮件：hzhou43@uic.edu

摘要

幂律分布广泛存在于化学物理学、地球物理学、生物学等领域。这些分布的自变量 x 有一个必需的下界，在许多情况下还有一个上界。从样本数据估计这些界限向来颇具难度，近期一种方法需要进行 $O(N^3)$ 次运算，其中 N 表示样本量。本文提出一种估计下界和上界的方法，只需进行 $O(N)$ 次运算。该方法的核心是计算由 N 个数据点组成的样本中最小的 x 个值和最大的 x 个值的均值，分别记为 $x\&_{\min}$ 和$x\&_{\max}$。将 $x\&_{\min}$ 或 $x\&_{\max}$ 拟合为 N 的函数，即可得出下界或上界的估计值。对合成数据的应用表明，该方法准确且可靠。

自然界和社会世界中的许多事件，其发生频率都遵循幂律分布：

$$p(x) = Ax^{-\alpha}, \qquad x > a, \tag{1}$$

其中，指数 α 通常介于 1 和 3 之间， A 是归一化常数，a 是下界。例子包括闪烁量子点的开关持续时间、[1,2]不同信使RNA的表达水平、[3,4] 不同国家每日新增的 COVID-19 病例数、[5]特定地区观测到的不同震级地震的发生频率，[6] 以及科学论文的引用次数。[7,8] 在许多情况下，自变量 x 的取值范围除了下界外，还有一个上界 b：

$$p(x) = Ax^{-\alpha}, \qquad a < x < b. \tag{2}$$

例子包括量子点的开启持续时间、[2] 每日 COVID-19 病例数、[5]以及特定地区地震的震级。[9] 人们已开发出许多方法，利用样本数据估计幂律分布的参数（α、 a和b）。虽然已有快速且稳健的方法可用于估计指数 α，但估计边界，尤其是上界 b，一向极具挑战性，也是本文的研究主题。

估计 α 最简单的方法是，在双对数图上将 $p(x)$ 与 x 之间的关系斜率作为估计值。更稳健的方法是采用最大似然估计。[10] 给定一个包含 N 个点 $\{x_i\}_{i=1,}$ 且被认为服从无上界幂律分布的样本，该估计值为

$$\alpha = 1 + N/\sum_{i=1}^{N} \ln(x_i/a). \tag{3}$$

请注意，此估计需要下界 a 的值。对 a 的估计则复杂得多。在 Clauset 等人提出的方法中，a 的估计值（记为 a&）从样本点 $\{x_i\}_{i=1,}$ 中选取，以使理论累积分布函数与经验累积分布函数（累积分布函数，CDFs）之间的柯尔莫哥洛夫–斯米尔诺夫距离最小：

$$KS = \max_{x \geq \hat{a}\&\in\{x_i\}_{i=1}^{N}} \left|\hat{C}(x) - C(x)\right|. \tag{4}$$

经验累积分布函数 $C@(x)$ 根据满足 $x_i \geq a\&$ 的样本点子集确定。理论累积分布函数为

$$C(x) = \int_a^x p(x')dx', \tag{5}$$

其中 a 和 α 设为当前估计值。这种方法很复杂，因为 α的估计与 a 的估计相互耦合。它涉及 $O(N^2)$ 次运算，因此成本也很高。

对于带有上界的幂律分布，情况会变得更糟。通常，人们会回避估计边界 a 和 b 的问题，转而分别将 a 和 b 设为样本中最小和最大的 data values。这种做法可能会导致对 α的估计产生显著偏差。有人提出了对边界进行修正的方法，将 $a\&$ 外推至小于最小 data value 的位置，并将 $b\mathrm{E}$ 外推至大于最大 data value 的位置。[9, 12, 13] 最近，Olmez 等人[14] 扩展了 Clauset 等人的方法，以同时估计 α、 a和 b 。本质上，他们将样本数据中的所有点对（例如满足 $x_i < x_j$的 x_i 和 x_j ）作为 a 和 b 的候选估计值进行检验（即 $a\& = x_i$ 和 $b\mathrm{E} = x_j$），并选择使 KS 最小的点对作为 $a\&$ 和 $b\mathrm{E}$ 的最终取值。该方法涉及 $O(N^3)$ 次运算。尽管如此，最终得到的 $a\&$ 仍不小于最小 data value，且 $b\mathrm{E}$ 仍不大于最大 data value。

在此，我提出一种估计下界和上界的新方法，每项涉及 $O(N)$ 次运算。该方法的核心是计算 N 点样本中最小的 x 和最大的 x 的均值，分别为 $x\&_{\min}$ 和$x\&_{\max}$。首先，我给出精确的 $x_{\min}$ 和 $x_{\max}$ 的表达式。[9, 10, 13] 首先请注意，累积分布函数 $C(x)$ 也表示数据点小于 x 的概率。同样，互补累积分布函数，

$$\tilde{C}(x) = 1 - C(x), \tag{6}$$

也是数据点大于 x 的概率。在包含 N 个点的样本中，某个特定点介于 x 和 $x + dx$ 之间，而其余 $N - 1$ 个点都大于 x 的概率为 $p(x)dx \cdot \mathrm{HCF}(x)\mathrm{I}^{-1}$。考虑到这 N 个点中的任意一个都可能是最小值，因此最小值为 x 的概率密度为

$$\pi(x) = Np(x)\left[\tilde{C}(x)\right]^{N-1}. \tag{7}$$

N 点样本中最小 x 的期望值或均值为

$$x_{\min} = \int_a^b x\pi(x)\,dx. \tag{8}$$

将 $p(x)$ 的表达式连同适当的归一化因子代入，公式（8）变为

$$x_{\min} = \frac{N(\alpha-1)}{[1-(b/a)^{-\alpha+1}]^N}\int_a^b (x/a)^{-\alpha+1}$$

$$[(x/a)^{-\alpha+1} - (b/a)^{-\alpha+1}]^{N-1}dx. \tag{9}$$

随着样本量 N 增加，$x_{\min}$ 应趋近于 a。这一趋势是本文方法估计 a 的基础。当上界为无穷大时，积分可计算得出

$$x_{\min} = \frac{N(\alpha-1)a}{N(\alpha-1)-1} \tag{10a}$$

$$= \frac{a}{1-B/N}, \tag{10b}$$

其中 $B = (\alpha-1)^{-1}$。当上界有限时，可将公式（9）转换为更适合数值积分的形式：

$$x_{\min} = \frac{Nb(\alpha-1)}{[(a/b)^{-\alpha+1}-]^N}\int_{a/b}^1 y^{-\alpha+1}(y^{-\alpha+1}-1)^{N-1}dy, \tag{11}$$

这里进行了代换 $y = x/b$。数值积分采用中点法实现，步长为 10^{-9}。

估计 a 的算法包括：（1）在一个 N 点的 t

样本；(2) 评估 M 个独立样本中的最小值均值（$x\&_{\min}$）；(3) 在保持 M 固定的同时，针对若干不同的样本量（N）重复上述步骤；以及 (4) 将 $x\&_{\min}$ 对 N 的依赖关系拟合为一个函数

$$\hat{x}_{\min} = \frac{\hat{a}}{1-(B/N)^{\gamma}}, \tag{12}$$

其中，我引入了一个额外参数 γ，以适应估计的 $x\&_{\min}$ 与精确 $x_{\min}$ 之间的偏差，以及上界（如果存在）可能产生的影响。

图1A展示了无上界以及具有 $\alpha = 2$ 和 $a = 1$时幂律的示例结果。通过随机数生成器，按照目标幂律生成了 $N = 10$、20、50、100、200、500 和 1000 个合成样本。在每个 N 下，$x\&_{\min}$ 取 $M = 100$ 个样本中最小值的均值；$x\&_{\min}$ 的误差估计为 $\mathrm{SD}/\sqrt{M}$，其中 SD 是 M 个最小值的标准差。$x\&_{\min}$ 与精确 $x_{\min}$ 之间的偏差大多在误差范围内。将 $x\&_{\min}$ 拟合到公式（12），得到 $a\& = 1.0009$，与真实值 1 极为接近。另外两个参数 B 和 γ 的值分别为 0.86 和 0.98，而两者的精确值均为 1。在 $\alpha = 1.5$ 和 2.5 时也得到了类似结果，拟合分别得到 $a\& = 1.002$ 和 1.0006。随着 α 增加，估计的a 与真实值之间的差异预计会减小，因为在较高的 α下，$x\&_{\min}$ 会更快地收敛到极限值 a 。

图1B展示了上界为 100 的幂律对应的结果。精确值 $x_{\min}$ [公式（11）] 略低于无上界幂律对应的值 [公式（10b）]。估计值 $x\&_{\min}$ 在误差范围内与精确值 $x_{\min}$ 一致。对公式（12）进行拟合后，在 $\alpha = 1.5$、2.0 和 2.5 处分别得到 $a\& = 1.002$、1.0009 和 0.9999。

接下来给出用于估计 b 的类似算法。对于一个包含 N 个点的样本，最大值为 x 的概率密度为

$$\Pi(x) = Np(x)[C(x)]^{N-1}. \tag{13}$$

在包含 N 个点的样本中，最大 x 的均值为

$$x_{\max} = \int_a^b x\Pi(x)\,dx \tag{14a}$$

$$= \frac{Na(\alpha-1)}{[1-(b/a)^{-\alpha+1}]^N}\int_1^{b/a} y^{-\alpha+1}(1-y^{-\alpha+1})^{N-1}dy\,, \tag{14b}$$

其中进行了替换 $y = x/a$ 。根据中点法，使用步长 10^{-5} 对方程 (14b) 进行了数值计算。随着 $x_{\max}$ 增大，b 应趋近于

N。图2A-C 展示了这一趋势，其中给出了在 b =100 和 1000 以及 α = 1.5、2.0 和 2.5 处计算的结果。为了估计 b，将 N 的两个数量级区间内的 $x_{\max}$ 值（例如 $[10^2, 10^4]$, ）拟合到

$$-\ln x_{\max} = -\ln \hat{b} + \frac{D}{1 + (N/E)^{\delta}}, \tag{15}$$

其中，bE、D、 E和δ 是自由参数。

所得的b 估计值汇总于图2D。首先来看真实值b 为100时的结果。当α = 1.5时，使用$x_{\max}$ 得到的估计值bE为100.5 ± 0.2，结果为$N \in [10, 10^3]$，因此，当在相对较低的N范围内进行估计时，bE与b 之间的差异已在1%以内。（“±”符号后的数字表示拟合参数的标准误。）当α = 2.0时，在$[10, 10^3]$的N 区间内，估计的bE为106 ± 1；与b 的差异在下一个更高的区间$[10^2, 10^4]$,内缩小至1%以内，此时bE为$= 100.24 \pm 0.08$。当 α = 2.5时， bE在$N \in [10, 10^3]$中为135 ± 5，在$N \in [10^2, 10^4]$中降至103.8 ± 0.9，最终在$[10^3, 10^5]$区间中达到1%的差异界限，降至100.16 ± 0.06。

当 b 高于 1000 时， $x_{\max}$ 需要更高区间内的 N 结果，才能将 bE 与 b 之间的差异控制在 1% 以内。在 α = 1.5 时，这一界限在 $N \in [10^2, 10^4]$ 中被突破。在 α = 2.0 时，bE 在 $[10^2, 10^4]$ 的 N 区间内为 1058 ± 13；与 b 的差异在下一个更高区间 $[10^3, 10^5]$, 中降至 1% 以内，此时 bE 为 1002.4 ± 0.8。在 α = 2.5 时， bE 在 $N \in [10^2, 10^4]$ 中为 1973 ± 149，在 $N \in [10^3, 10^5]$ 中降至 1126 ± 24，最终在 $[10^4, 10^6]$ 的 N 区间内突破 1 % 的差异界限，达到 1009 ± 2。

从样本数据估计b 的算法包括：(1) 找出一个含N个点的样本中的最大值；(2) 计算M个独立样本中最大值的均值（x&max）；(3) 在保持M 不变的情况下，对跨越两个数量级区间的样本量（N）重复上述步骤；(4) 将x&$_{\max}$ 对N 的依赖关系拟合到公式（15）。

图3A-C展示了从样本数据估计b 的示例结果，其真实值为100，且α = 1.5、2.0和2.5。同样，合成数据由随机数生成器生成，遵循目标幂律；$x\&_{\max}$ 则估计为M = 100个样本中最大值的均值。样本量为10至10^5 的样本所得$x\&_{\max}$ 值与精确的$x_{\max}$ 相符，误差范围内一致。因此，通过将$x\&_{\max}$ 拟合到公式（15）来进一步估计b ，所得结果与使用精确$x_{\max}$时相近。区别之一是，在N 跨越两个数量级的区间内，需要更多的$x\&_{\max}$ 值（例如，图3A-C中为23个，而图2A-C中为7个），才能确保bE拟合值的可靠性。bE值汇总于图4D。与真实b 的差异与使用精确$x_{\max}$ 值时发现的差异相似（图3D）。在α = 1.5时，使用$N \in [10, 10^3]$中的$x\&_{\max}$ 值，估计的bE为100.6 ± 0.8，误差为< 1%。在α = 2.0时，估计的bE在$[10, 10^3]$的N 区间内为111 ± 6；在下一个更高的区间$[10^2, 10^4]$，中，bE为99.4 ± 0.4，与真实b 的差异缩小至1%以内。在α= 2.5时，N 区间的$[10, 10^3]$未能得出可靠的bE拟合值。在稍高的区间$[20, 2 \times 10^3]$中，拟合得到bE = 139 ± 36，但标准误仍然较高。在下一个N 区间$[10^2, 10^4]$中，bE为97 ± 1，与真实b 的差异和标准误均较小。最后，在$[10^3, 10^5]$区间内，bE跨过1%的差异界限，达到100.1 ± 0.3。

总之，本文提出了一种简单的 $O(N)$ 算法，用于估计幂律分布的下界和上界。该算法通过获取样本数据中的最小值和最大值，并将这些值拟合为样本量的函数来实现。下界和上界的估计与指数的估计相互独立，且无需输入任何指数相关信息。估计得到的边界可能超出观测数据的范围。最后，该算法可以轻松扩展到其他类型的分布函数。

致谢

本文　这项工作由美国国立卫生研究院的GM118091号资助支持　　.

数据可用性

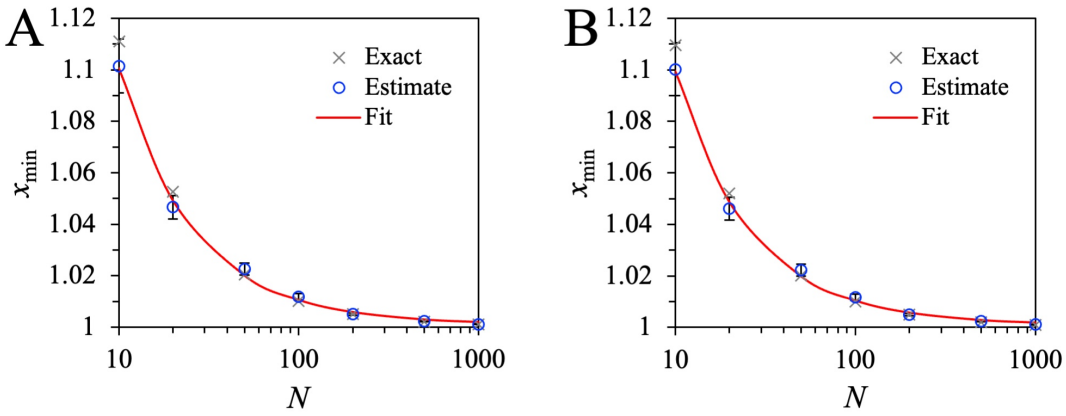
支持本研究发现的数据可在提出合理请求后向通讯作者获取。

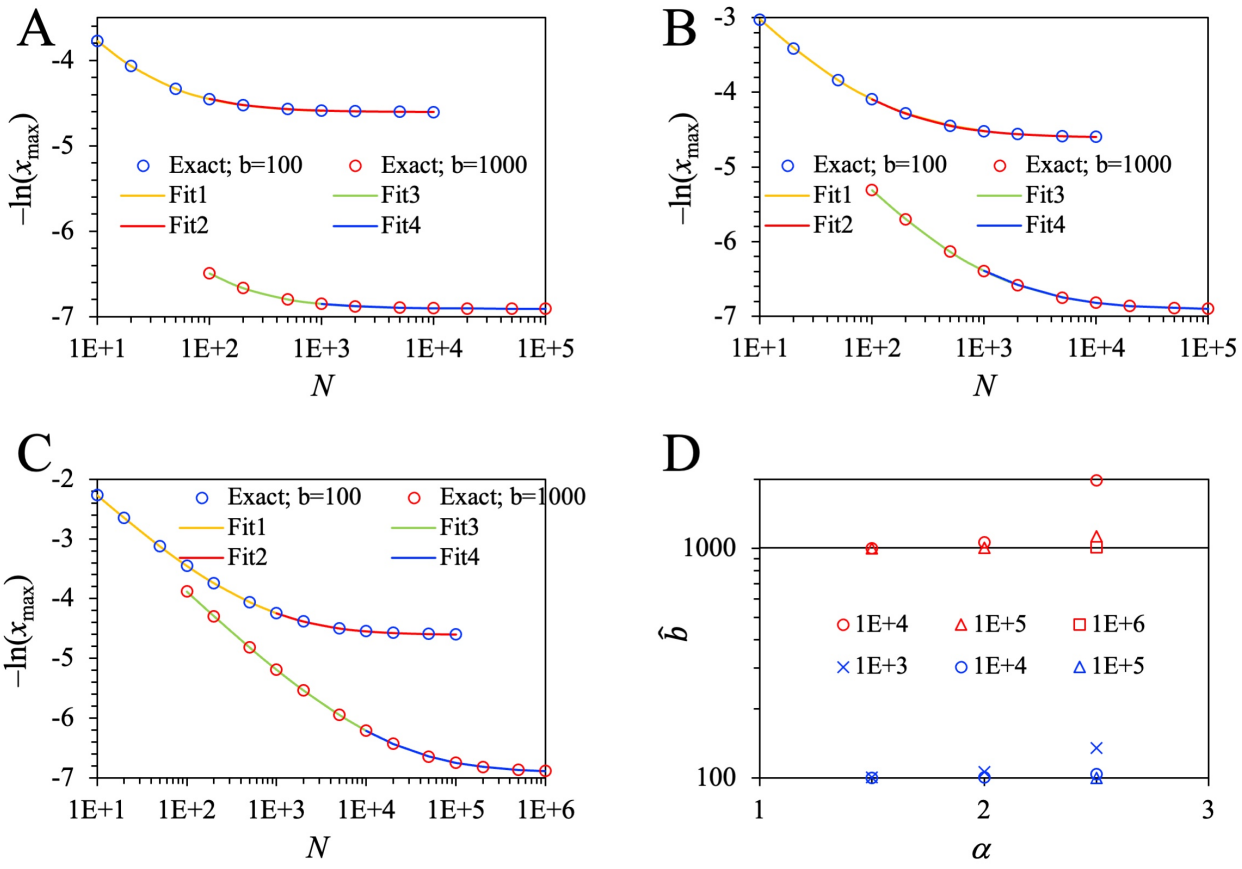
参考文献

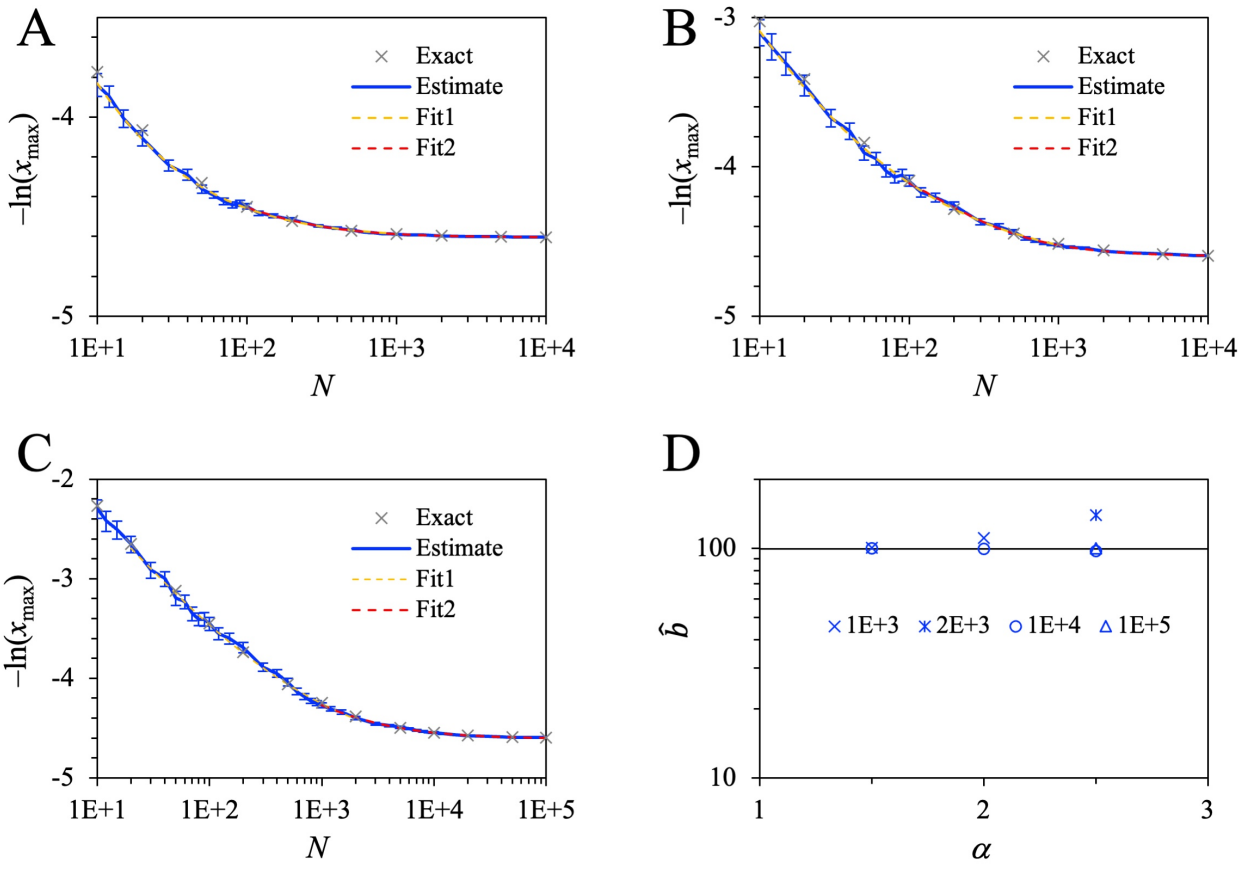
[1] M. Kuno、D. P. Fromm、H. F. Hamann、A. Gallagher 和 D. J. Nesbitt，《化学物理杂志》**115**，1028（2001）。[2] K. T. Shimizu、R. G. Neuhauser、C. A. Leatherdale、S. A. Empedocles、W. K. Woo 和 M. G. Bawendi，《物理评论B辑》**63**，205316（2001）。[3] C. Furusawa 和 K. Kaneko，《物理评论快报》**90**，088102（2003）。[4] H. R. Ueda、S. Hayashi、S. Matsuyama、T. Yomo、S. Hashimoto、S. A. Kay、J. B. Hogenesch 和 M. Iino，《美国国家科学院院刊》**101**，3765（2004）。[5] B. Blasius，《混沌》**30**，093123（2020）。[6] B. Gutenberg 和 C. F. Richter，《美国地震学会通报》**34**，185（1944）。[7] D. J. d. S. Price，《科学》**149**，510（1965）。[8] G. J. Peterson、S. Pressé 和 K. A. Dill，《美国国家科学院院刊》**107**，16023（2010）。[9] A. Kijko 和 M. Singh，《地球物理学报》**59**，674（2011）。[10] M. E. J. Newman，《当代物理学》**46**，323（2005）。[11] A. Clauset、C. R. Shalizi 和 M. E. J. Newman，《SIAM评论》**51**，661（2009）。[12] T. Maschberger 和 P. Kroupa，《皇家天文学会月刊》**395**，931（2009）。[13] J. Zhang，《统计学》**47**，792（2013）。[14] F. Olmez、P. R. Kramer、J. Fricks、D. R. Schmidt 和 J. Best，《统计计算与模拟杂志》**91**，1524（2021）。

图注

图 1. $x\&_{\min}$ 与精确的$x_{\min}$ 进行比较，并拟合至公式（12），以估算下界a。(A) 无上界的幂律分布；$\alpha = 2$ 和$a = 1$。(B) 存在上界$b = 100$的幂律分布对应的结果。图 2. 将精确的$x_{\max}$ 拟合至公式（15），以估算上界b。(A) α =1.5。拟合1、拟合2、拟合3和拟合4分别使用N以下范围的数据：$[10, 10^3]$, $[10^2, 10^4]$, $[10^2,10^4]$, 和$[10^3, 10^5]$, 。(B) $\alpha = 2$。拟合1、拟合2、拟合3和拟合4分别使用N以下范围的数据：$[10, 10^3]$, $[10^2, 10^4]$, $[10^2, 10^4]$, 和$[10^3, 10^5]$, 。(C) α =2.5。拟合1、拟合2、拟合3和拟合4分别使用N以下范围的数据：$[10, 10^3]$, $[10^3, 10^5]$, $[10^2,10^4]$, 和$[10^4, 10^6]$, 。(D) 估计的b。图例中标出了用于估算b 的两个数量级N 区间的上限。图 3. 将$x\&_{\max}$ 拟合至公式（15），以估算上界b。真实的 $b = 100$。(A) $\alpha = 1.5$。拟合1和拟合2分别使用N 以下范围的$[10, 10^3]$ 和$[10^2, 10^4]$, 数据。(B) $\alpha = 2$。拟合1和拟合2分别使用N 以下范围的$[10, 10^3]$ 和$[10^2, 10^4]$, 数据。(C)$\alpha = 2.5$。拟合1和拟合2分别使用N 以下范围的$[20, 2 \times 10^3]$ 和$[10^3, 10^5]$, 数据。(D) 估计的b。图例中标出了用于估算b 的两个数量级N 区间的上限。


A
B
1.12
1.1
1.08
1.06
1.04
1.02
1
10
100
1000
N
x_{min}
Exact
Estimate
Fit


A
B
C
D
$-\ln(x_{max})$
N
Exact; b=100
Exact; b=1000
Fit1
Fit2
Fit3
Fit4
1E+1
1E+2
1E+3
1E+4
1E+5
1E+6
$\hat{b}$
α
1000
100
1
2
3
1E+4
1E+5
1E+6
1E+3
1E+4
1E+5


A
B
C
D
Exact
Estimate
Fit1
Fit2
$-\ln(x_{max})$
N
1E+1
1E+2
1E+3
1E+4
1E+5
-2
-3
-4
-5
$\hat{b}$
100
10
1
2
3
α
1E+3
2E+3
1E+4
1E+5