

用 Grassberger-Procaccia 方法 计算吸引子维数的基本限制*

洪 时 中

成都市地震局, 成都 610015

洪 时 明

成都科技大学高压物理研究所, 成都 610065

1993 年 8 月 11 日收到

使用 Grassberger-Procaccia 方法(简称 G-P 方法)计算吸引子的维数时, 必然受到数据点数 N 的限制. Smith (1988 年), Eckmann 与 Ruelle (1992 年) 曾先后研究过这种限制. 本文指出了这些研究存在的问题, 另外推导出了新的关系式 $N_{\min} > \sqrt{2} \cdot (\sqrt{27.5})^D$. 这一要求比 Smith 给出的 $N_{\min} \geq 42^M$ 大为降低, 可以拓展 G-P 方法的使用范围.

PACC: 0545

一、引 言

Grassberger 与 Procaccia^[1] 在 Pakard 等人及 Takens 重建相空间方法的基础上, 提出了一种由观测得到的标量时间序列计算其动力系统吸引子关联维数 D_2 的方法, 即 G-P 方法. 这一算法给出了由实验数据或观测数据探索其动力系数非线性特征的新途径, 受到了许多应用领域的重视. 在很短的时间内, 就有采用 G-P 方法研究气候变迁、脑电波、地震活动、经济现象等的大量文献问世. 但是, 人们很快发现, G-P 方法的使用受到数据点个数(即标量时间序列的长度) N 的限制, 因而对采用 G-P 方法计算所得结果的可靠性提出了种种质疑. 特别是 Smith^[2] 给出了 $N_{\min} \geq 42^M$ 的判据 ($N_{\min}$ 为数据点个数的下限, M 为维数的整数部分)之后, 更使人感到 G-P 方法的应用前景十分悲观. 因此, 弄清样本容量对 G-P 方法究竟有多大的限制, G-P 方法实际应用的可能性究竟有多大? 就具有十分重要的意义.

二、E-R 关系式存在的问题

Eckmann 与 Ruelle^[3] 在最初并不知道 Smith 工作的情况下, 对维数计算的基本限制独立地进行了研究. 他们根据

* 国家地震局“八五”攻关项目资助的课题.

$$\eta(r) = \text{const} \times r^D, \quad (1)$$

$$\begin{aligned} \eta(r) &= \text{const} \times L^D \\ &\approx \frac{1}{2} N^2, \end{aligned} \quad (2)$$

$$\eta(r) \gg 1, \quad (3)$$

经过简单的代数变换得到

$$d_{\max} = \frac{2 \log N}{\log(1/\rho)}, \quad (4)$$

其中 r 为距离, $\eta(r)$ 为相空间中距离在 r 以内的不重复点对数, L 为吸引子的“直径”, 亦即该吸引子中点对距离的最大值, N 为数据点的总数, D 为吸引子的维数, $d_{\max}$ 为所允许计算的最大维数, $\rho = r/L$, 他们建议取 $\rho = 0.1$.

为便于比较, 将(4)式改写为

$$N_{\min} \geq (\sqrt{1/\rho})^D, \quad (5)$$

$N_{\min}$ 为所必须的最小的数据点个数。当 $\rho = 0.1$ 时, (5)式变为

$$N_{\min} \geq (\sqrt{10})^D. \quad (6)$$

这一要求远低于 Smith, 要计算一个 4 维的吸引子, 只需要 100 个数即够。

我们发现, Eckmann 与 Ruelle 所采用的前提并不完全正确, (2)式就不能成立。因为 $\log r - \log C(r)$ ($C(r)$ 为关联积分) 曲线的上部(即当 r 接近 L 时)必然会出现弯曲, 而(2)式则假定 $\log r - \log C(r)$ 的线性关系一直延伸至 $r = L$ 处, 将该曲线上部的弯曲部分完全当作直线来处理, 这是不符合实际的。计算表明, 当将曲线与直线的偏离控制在 10% 之内时, 弯曲部分将占去约 $2^2 - 2^3$ 的宽度, 这时如果仍取 $\rho = 0.1$, 剩下的线性部分(即“无标度区”)已经几乎不复存在。所以(4), (5)和(6)式的要求大大偏低, 用它们来衡量数据点个数是否足够是没有什么意义的。

此外, 还应当指出, 文献[3]在推导过程中因疏忽还漏掉了一个系数 2, (4)至(6)式应分别改为

$$d_{\max} = \frac{2 \log N - \log 2}{\log(1/\rho)}, \quad (4')$$

$$N_{\min} \geq \sqrt{2} \cdot (\sqrt{1/\rho})^D, \quad (5')$$

$$N_{\min} \geq \sqrt{2} \cdot (\sqrt{10})^D. \quad (6')$$

三、Smith 关系式存在的问题

Smith^[2] 较好地研究了 $\log r - \log C(r)$ 曲线上部的弯曲, 他利用直线段上随机选择的两个点的距离小于 l 的几何型概率模型导出

$$C(l) = [l(2 - l)]^M, \quad (7)$$

其中 l 为相空间中的距离, $C(l)$ 为关联积分, M 为维数。其斜率 $\nu(l)$ 可对(7)式求导得出

$$\nu(l) = M[1 - l/(2 - l)]. \quad (8)$$

令 $l_{\max}$ 和 $l_{\min}$ 分别为“无标度区”的上限和下限, Q 为质量系数, 即 ν 与其真值 M 之比, 且 $0 \leq Q \leq 1$, $R = l_{\max}/l_{\min}$, 即“无标度区”的宽度. 将它们代入(8)式整理后得

$$l_{\min} = \frac{2(1-Q)}{R(2-Q)}. \quad (9)$$

Smith 认为, 在将相空间分割为边长为 $l_{\min}$ 的 M 维超立方体时要求

$$N_{\min} \geq (1/l_{\min})^M \nu. \quad (10)$$

由(9)式与(10)式得

$$N_{\min} \geq \left[\frac{R(2-Q)}{2(1-Q)} \right]^M. \quad (11)$$

文献[2]中建议取 $R = 4$, $Q = 0.95$, 此时

$$N_{\min} \geq 42^M. \quad (12)$$

这是一个相当高的要求, 当 $M = 3$ 时, 要求观测数据至少要有 74088 个点; $M = 6$ 时, 要有 5,489,031,774 个点. 实际资料很难达到这一要求, 因而使得 G-P 方法的应用受到极大的限制.

Smith 的工作并非无懈可击. 我们认为, 十分关键的(10)式的要求就不尽合理.

(10) 式意味着每一个边长为 $l_{\min}$ 的 M 维超立方体都平均至少有一个点, 这并不是 G-P 方法本身的要求. G-P 方法不是数盒子法, 它不需要将空间划分为盒子(超立方体), 更不要求每个盒子都有点. 对一些已知其动力方程的混沌吸引子所作的数字试验也与(10)式大相径庭. 根据 Grassberger 与 Procaccia^[4]所作的计算结果, 对于 Hénon 映射, 当 $N = 15000$ 时, $l_{\min}$ 至少可达 2^{-22} (归一化值, 下同), 平均每个超立方体仅 0.0036 个点. 对于 Lorenz 方程, $N = 15000$ 时, $l_{\min}$ 可达 2^{-13} , 平均每个超立方体仅 0.0002 个点, 而在这两个例子中, $\log r - \log C(r)$ 的线性关系仍然很好, “无标度区”的宽度可以分别达到 2^{20} 和 2^{10} . 由此可见, (10) 式的要求的确是不合理的. 正是由于这一要求大大偏高, 才导致(11), (12)式的要求也大大偏高.

四、维数计算基本限制的新公式

Smith 所采用的模型认为: 在相空间中点的分布是完全随机的、相互独立的. 实际混沌动力系统的吸引子当然并非如此, 但这一模型可以较好地描述 $\log r - \log C(r)$ 曲线顶端的弯曲, 与实际的 $\log r - \log C(r)$ 曲线相当接近(图 1), 不失为一种相当好的近似, 因而是可以接受的.

另一方面, Eckmann 与 Ruelle 提出的(3)式才是 G-P 方法的基本要求. 由(3)式可得

$$\eta(l_{\min}) > 1. \quad (13)$$

我们主张, 采用 Smith 的模型, 同时采用(13)式取代(10)式, 即可推出新的维数计算

1) 文献[2]中此式有印刷错误.

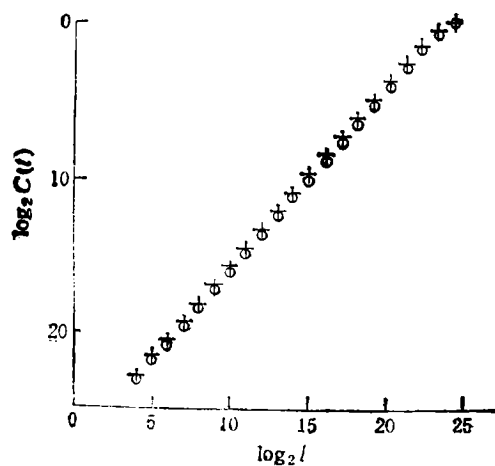


图1 Hénon 映射的关联积分曲线^[9] ($N = 15000$, $a = 1.4$, $b = 0.03$, 用○表示)与几何型概率模型(即(7)式)计算的结果(用+号表示)

基本限制的公式。

当 l 很小时, (7)式变为

$$C(l) \approx (2l)^M, \quad (14)$$

同时还有

$$C(l) \approx \frac{\eta(l)}{\frac{1}{2} N^2}. \quad (15)$$

令 $M = D$, 由(9), (13)至(15)式可得

$$N_{\min} > \sqrt{2} \cdot \left[\sqrt{\frac{R(2-Q)}{4(1-Q)}} \right]^D. \quad (16)$$

这就是我们修正的结果。若仍取 $R = 4$, $Q = 0.95$, 则

$$N_{\min} > \sqrt{2} \cdot (\sqrt{21})^D. \quad (17)$$

表1 Smith 与本文所得 $N_{\min}$ 值的比较

维 数 D	Smith 给出的 $N_{\min}$ $R = 4$ $Q = 0.95$	本文给出的 $N_{\min}$	
		$R = 4$ $Q = 0.95$	$R = 10$ $Q = 0.90$
1	42	7	8
2	1764	30	39
3	74088	137	204
4	3,111,696	624	1070
5	130,691,232	2859	5609
6	5,489,031,744	13,098	29,412

但我们认为 $R = 4$ 要求太低,建议取 $R = 10$ (即“无标度区”的宽度为一个数量级);同时,考虑到实际求 ν 时不是靠求导数,而是用最小二乘法,因此不必取 $Q = 0.95$, 建议取 $Q = 0.90$ 。这时,(16)式变为

$$N_{\min} > \sqrt{2} \cdot (\sqrt{27.5})^D. \quad (18)$$

我们推荐用(18)式来衡量用 G-P 方法计算时数据点是否足够。表 1 比较了 Smith 与我们工作的结果。从表 1 中可以看出,(18)式的要求远比 Smith 的为低,在低维情况下,许多实际资料是可以满足(18)式的。

五、讨 论

1. 上述工作都只研究了系统误差,没有涉及到随机误差问题。Theiler^[5] 指出,在用 G-P 方法计算关联维数时,其误差一般与 $O(1/\sqrt{N})$ 甚至 $O(1/N)$ 成正比; Ben-Mizrachi 等人^[6]指出,随机误差主要影响 r 较小的一端,它造成 $\log r - \log C(r)$ 曲线下端的弯曲。一般来说,在 G-P 方法计算中,系统误差比随机误差更为重要。

2 应当指出,本文给出的判据只是判断数据点数是否足够的必要条件,而不是其充要条件。也就是说,如果不满足(18)式,就可以否定 G-P 方法计算的结果;但是,如果仅仅满足(18)式,并不能证明数据点数就已经足够。那么,究竟 N 要多大才够呢? 我们认为,这个问题说到底就是要判断是否存在足够宽的“无标度区”。在满足(18)式的前提下,如果所得出的 $\log r - \log C(r)$ 曲线有足够宽的“无标度区”,则说明数据点数是足够的,噪声的影响也可以克服;否则,可能是数据点数仍然不够,也可能是噪声太大,甚至可能该动力系统根本不是确定性的,这时 G-P 方法已无法使用。对于“无标度区”的客观判定,笔者已提出了一种“自相似比法”^[7],可望较好地加以解决。

3. 尽管我们给出的判据仍然只是一个必要条件,尽管我们所采用的 Smith 模型只是一种近似,尽管本文没有考虑随机误差的影响,但本文所给出的结果仍然十分重要。一方面,它打破了 G-P 方法应用中某些人为的禁区,拓宽了 G-P 方法的应用范围;另一方面,它可以使我们对 G-P 方法计算结果的可靠性作一初步判断,提供了一种实用的判据。

朱照宣、李后强、黄登仕、艾南山、王文均、傅军、郑兆蕊等同志与笔者进行了有益的讨论,谨此致谢。

[1] P. Grassberger and I. Procaccia, *Phys. Rev. Lett.*, **50**(1983), 346.

[2] L. A. Smith, *Phys. Lett.*, **133**(1988), 283.

[3] J. P. Eckmann and D. Ruelle, *Physica*, **D56**(1992), 185.

[4] P. Grassberger and I. Procaccia, *Physica*, **D9**(1983), 189.

[5] J. Theiler, *Measures of Complexity and Chaos*, edited by N. B. Abraham *et al.*, (1991), p.199-202.

[6] A. Ben-Mizrachi *et al.*, *Phys. Rev.* **A29**(1984), 975.

[7] 洪时中、洪时明,大自然探索,**12**(1993),2;53.

FUNDAMENTAL LIMITS ON ESTIMATING DIMENSION OF ATTRACTOR USING GRASSBERGER-PROCACCIA ALGORITHM

HONG SHI-ZHONG

Seismological Bureau of Chengdu, Chengdu 610015

HONG SHI-MING

*Chengdu University of Science and
Technology, Chengdu 610065*

(Received 11 August 1993)

ABSTRACT

Fundamental limits on the minimum size $N_{\min}$ of data set is an important problem in estimating the dimension D of attractor using Grassberger-Procaccia algorithm. We modify the results of Smith (1988), Eckmann and Ruelle (1992), and suggest a new criterion $N_{\min} > \sqrt{2} \cdot (\sqrt{27.5})^D$.

PACC: 0545