基于概率密度分布型变化的突变检测新途径*

成海英[1)] 何文平[2)†] 何涛[3)] 吴琼[4)]

1) (盐城工学院基础教学部, 盐城 224051)

2) (国家气候中心, 北京 100081)

3) (济南市环境监测中心站, 济南 250014)

4) (国家卫星气象中心, 北京 100081)

(2011 年 4 月 9 日收到; 2011 年 5 月 18 日收到修改稿)

对于一个稳定的动力系统而言, 系统变量的概率密度具有较为稳定的分布型, 而当系统的动力学结构发生变化后可能会导致系统变量的分布型发生不同程度的变化. 鉴于此, 本文从识别系统变量的概率密度分布的微小变化角度出发, 将描述时间序列概率分布特征的偏度系数和峰度系数应用于时间序列的突变检测中. 数值试验结果表明, 偏度系数和峰度系数对突变信号具有很好的识别能力, 进而揭示了一条检测突变的新途径. 进一步的研究表明, 新方法的检测结果对于子序列长度的选择具有较小的依赖性.

关键词: 偏度系数, 峰度系数, 概率密度分布, 突变检测

PACS: 92. 60. Wc

1 引 言

经典的牛顿力学的研究对象是连续的、渐变的、光滑变化的现象. 譬如, 地球围绕太阳旋转, 有规律地周而复始地连续不断进行, 使得人们能够采用经典的微积分精确地预测其未来的运动状态. 但是自然界许多现象并非都是连续的、渐变的过程, 还存在许多不连续的、突然的变化. 例如, 西风急流的跳跃性北移 [1], 随着印度季风的爆发, 赤道印度洋上空区域平均涡动动能的突然增长, 印度洋季风爆发前平均涡动动能为 20 m^2/s^2, 而季风爆发后迅速增大至 80 m^2/s^2[2]; 在医学上, 系统会发生自发性的故障, 如哮喘发作或癫痫发作; 在全球金融领域, 如股市、楼市的崩溃; 地球系统中海洋环流或气候的突变; 牧场、鱼群数量或野生动物数量的灾难性变化; 火山的突然爆发, 地震的突然发生、房屋的突然倒塌等 [3−5]. 所幸的是, 人们已经逐渐认识到包括生态系统、生物体、机械信号、金融市场、气候等在内的很多复杂动力系统中广泛存在着突变现象, 即系统在一定的条件下会从一种相对稳定的状态转变到另一种相对稳定的状态. 因此, 对突变的检测就显得尤为重要.

长期以来, 国内外学者已就这一问题开展了大量的研究, 如传统的突变检测方法 —— 滑动 t- 检验、Cramer 法、Yamamoto 信噪比方法以及 M-K 法等 [6−8]; Livina 等 [9] 和何文平等 [10,11] 基于时间序列的长期记忆性特征, 先后提出了时间序列突变检测的多种新方法. 近年来, 国内已有学者发展了一系列新的突变检测方法, 如条件熵、动力学自相关因子指数、高阶矩等 [12−16]. 王启光和张增平 [17]、何文平等 [18] 则基于度量时间序列复杂性程度的近似熵方法相继发展了两种用于突变检测的新技术. 此外, 还有学者将小波分析、粒子滤波等方法用于突变检测中 [19,20]. 这些研究工作极大地丰富了突变检测方法, 提供了检测突变的多种不同途径. 但是任何一种方法都不是万能的, 具有一定的适用范围, 因此, 需要进一步针对不同的突变检测问题, 发展相应的突变检测新技术.

对于一个稳定的动力系统而言, 系统变量的概率密度具有较为稳定的分布型, 而当动力学结构发

* 国家自然科学基金 (批准号: 40905034, 41175067, 40930952) 和公益性行业 (气象) 科研专项 (批准号: GYHY201106015, GYHY201106016) 资助的课题.

† E-mail: wenping_he@163.com

生变化后可能会导致系统变量的分布型发生不同程度的变化. 鉴于此, 本文另辟蹊径, 从系统变量的概率密度分布的微小变化角度出发, 捕捉和识别时间序列中的突变信号, 为突变检测提供一条新的途径.

2 方 法

本文主要侧重于研究如何将描述时间序列概率分布特征的偏度系数和峰度系数应用于时间序列的突变检测中. 偏度系数通常用于定量描述要素的概率密度分布与正态分布的差异, 峰度系数则是用来度量数据在中心聚集程度的物理量, 它反映了概率密度分布曲线顶端尖峭或扁平的程度. 有时两组数据的算术平均数、标准差和偏度系数都相同, 但它们分布曲线顶端的高耸程度却不同.

2.1 偏度系数和峰度系数

对于时间序列$\{x_i, i = 1, N\}$, 其偏度系数 g 可以写为

$$g = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)\sigma^3}, \tag{1}$$

式中, σ 为样本的标准偏差, 当概率分布密度曲线为正态分布时, 偏度系数为 0; 当偏度 g <0 时, 分布具有负偏离, 也称左偏态, 此时均值位于峰值的左边; 当 g >0 时, 情况正好相反.

峰度系数是描述分布形态的陡缓程度的物理量, 是一个表征概率密度分布曲线在平均值处峰值高低的特征数. 统计上是用四阶中心矩来测定峰度的, 实际工作常利用四阶中心矩与标准偏差四次方的比值作为衡量峰度的指标, 通常可以写为

$$k = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)\sigma^4}, \tag{2}$$

当峰度系数 k = 3 时表示该分布为正态分布; $k > 3$ 表示比正态分布陡峭; $k < 3$ 表示比正态分布平坦.

2.2 基于概率密度分布型变化的突变检测方法

对于一个稳定的动力系统而言, 系统变量的概率密度具有较为稳定的分布型, 而当系统的动力学结构发生变化后可能会导致系统变量的分布型发生不同程度的变化. 鉴于此, 本文从识别系统变量的概率密度分布的微小变化角度出发, 将描述时间序列概率分布特征的偏度系数和峰度系数应用于时间序列的突变检测中. 基于偏度系数和峰度系数的突变检测方法其主要步骤类似于滑动去趋势波动分析 [10]. 具体如下: 首先选择一定长度的子序列, 计算该子序列的偏度系数和峰度系数; 随后逐步移动该子序列, 但保持其长度不变; 再次计算新子序列的偏度系数和峰度系数. 如此重复操作, 直至子序列移动至所分析序列的末端. 最后, 基于偏度系数和峰度系数随子序列长度的演变特征来判别是否有突变发生.

3 数值试验

本文主要以 Logistic 模型 [21] 为例, 研究在动力系统的参数和动力学方程形式发生突变两种不同情形下, 基于概率密度分布型的微小变化, 考察偏度系数和峰度系数对于时间序列中突变信号的捕捉与识别能力. 经典的 Logistic 模型可写为

$$x_{n+1} = ux_n(1 - x_n), \quad x \in [0, 1], \tag{3}$$

式中 x_n 为第 n 代种群数, x_{n+1} 为第 n+1 代种群数; u 是一个大于 0 且小于 4 的控制参数, 当 3.569945672 < u < 4.0, 系统进入混沌状态. 在数值试验中, 首先考虑 Logistic 模型发生参数突变的情形, 即其控制参数在某一时刻由 3.8 突然减小为 3.7. 基于此, 图 1 给出了 Logistic 模型在 n = 10001 时参数 u 由 3.8 突然减小为 3.7 时的理想突变时间序列, 该序列总长度为 20000. 由于参数 u 减小后种群的方差随之减小, 造成图 1(a) 中的突变点过于明显, 为了避免这一缺陷, 分别将突变前后的数据进行标准化, 即将变量值减去其平均值, 然后除以该变量的标准差, 标准化公式为

$$y_i = \frac{x_i - \overline{x}_1}{\sigma_1}, \quad i = 1, 2, \cdots, 10000, \tag{4}$$

$$y_i = \frac{x_i - \overline{x}_2}{\sigma_2}, \quad i = 10001, 10002, \cdots, 20000, \tag{5}$$

在 (4) 和 (5) 式中, $\overline{x}_1$ 和 $\overline{x}_2$ 分别为参数 u 改变前后样本的平均值,σ_1,σ_2 为参数 u 改变前后样本的标准偏差. 经过标准化后的理想时间序列 (IS1) 已在图 1(b) 中给出, 其均值为 0, 标准方差为 1. 从图中可以看出, 标准化后的时间序列很难不通过突变检测工具来识别出其中的突变点, 而且标准化处理基

本不会影响数据的概率密度分布特征.

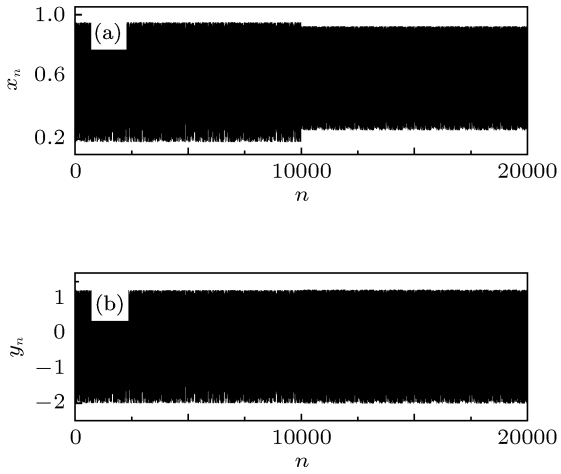

图 1 Logistic 模型发生参数突变的情形 (a) 在 n=10001 时, 参数 μ 由 3.8 突然减小为 3.7, 时间序列总长度为 20000; (b) 经过标准化后的理想时间序列 (IS1)

偏度系数和峰度系数对理想序列 IS1 的检测结果已由图 2 给出. 由于 Logistic 模型的偏度系数明显小于 0, 根据偏态指数的物理意义可知,Logistic 模型的概率密度分布型呈现左偏态. 当子序列长度取 100 时, 我们发现偏度系数的演化可以分为两个明显不同的阶段 (图 2(a)): 在 n <10001 前种群数演变的偏度系数总体上明显大于 n >10000 以后, 这表明自 n=10000 前后, 种群的概率密度分布形态发生了显著改变, 即相对于正态分布而言, 左偏的程度越发明显. 为了进一步研究偏度系数的检测结果的稳定性, 我们考察了不同子序列长度下偏度系数对 IS1 的检测结果. 图 2(b) 和 (c) 分别给出了子序列长度为 200 和 500 时的情形, 从中不难看出偏度系数的演化类似于子序列长度为 100 时的情形, 即大致以 n=10000 为界, 偏度系数可分为两个明显不同的演化阶段, 而且随着子序列样本量的增大, 这种差异越发显著. 所不同的是随着子序列长度的增加, 偏度系数的方差随之减小. 究其原因, 主要在于随着样本量的

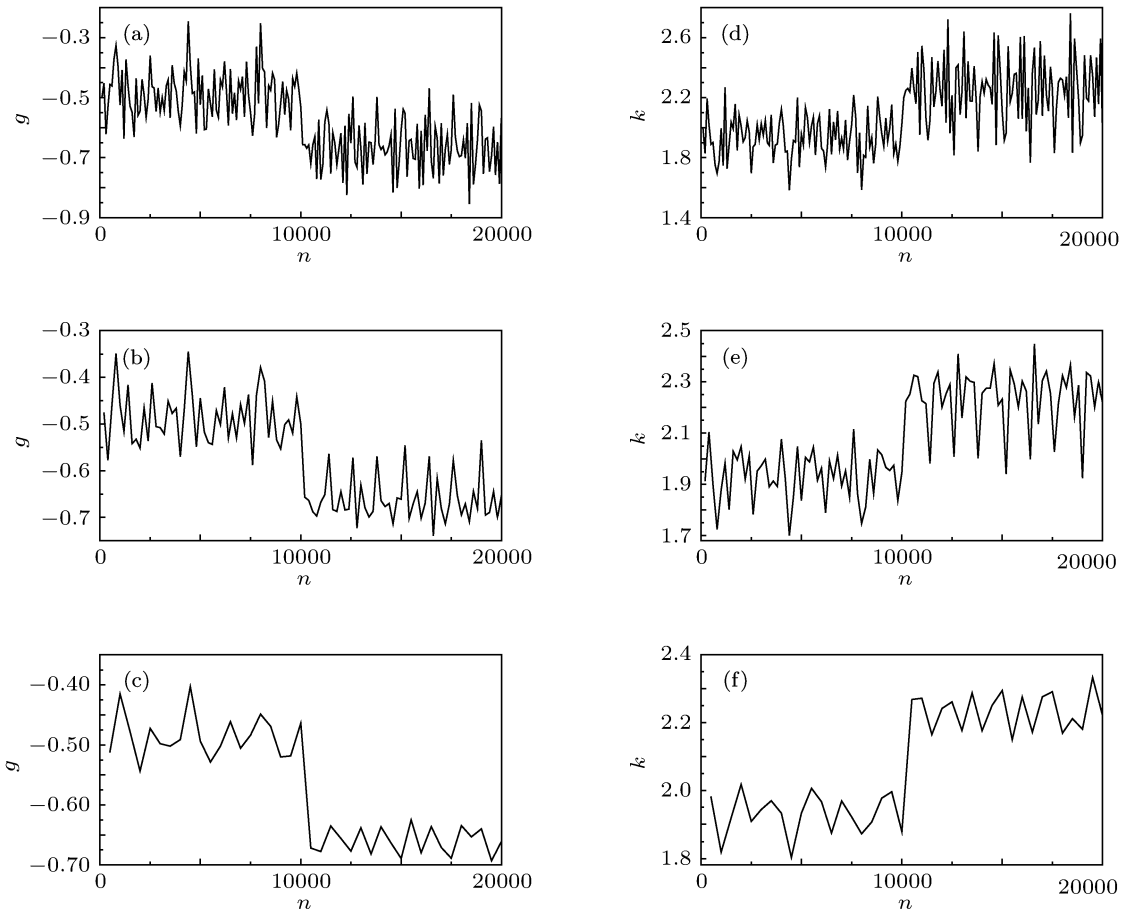

图 2 偏度系数和峰度系数对理想系列 IS1 的突变检测 (a), (b), (c) 分别为子序列长度为 100, 200, 500 时偏度系数的检测结果;(d), (e), (f) 分别为子序列长度为 100, 200, 500 时峰度系数的检测结果

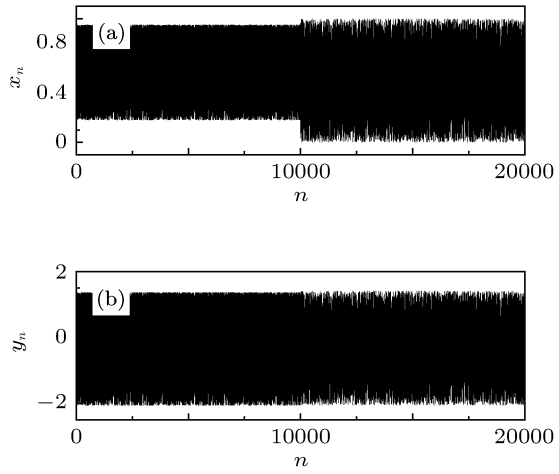

图 3 种群的演化发生动力学结构突变 (a) 在 n = 10001 时, 种群的演化由 Logistic 模型突变变为随机行动, 时间序列总长度为 20000; (b) 经过标准化后的理想时间序列 (IS2)

增加, 偏度系数的计算更加接近真实值, 减小了由于样本量过少所引起偏度系数波动过大的现象. 试验中, 对不同长度的子序列进行了大量对比分析, 发现检测结果类似于子序列长度取 100, 200 和 500 时的情形.

由于偏度系数主要描述的是要素的概率密度分布相对于正态分布时的偏离程度, 而实际中有时两组数据的算术平均数、标准差和偏度系数都相同, 但它们的分布曲线顶端的高耸程度却不同. 因而偏度系数仅仅为反映出概率密度分布特征的物理量之一, 显然, 当要素的偏度系数未发生变化而概率密度分布曲线的陡峭程度却发生变化时, 单靠偏度系数不能对此进行有效识别. 鉴于此, 本文中将峰度系数应用于要素的突变检测中.

从理想时间序列 IS1 的峰度系数检测结果来看, 在不同子序列长度下得到的峰度系数 k 均小于 3 (图 2(d)—(f)), 表明理想时间序列较正态分布平坦, 类似于偏度系数的检测结果. 峰度系数的演变具有一个共同的特点, 即 n = 10000 前后, 峰度系数发生了一次明显的均值突变 (图略), 曲线的峰度更加接近正态分布, 而且随着子序列长度的增大, 即样本量的增多, 导致峰度系数在 n = 10000 前后的差异更加明显.

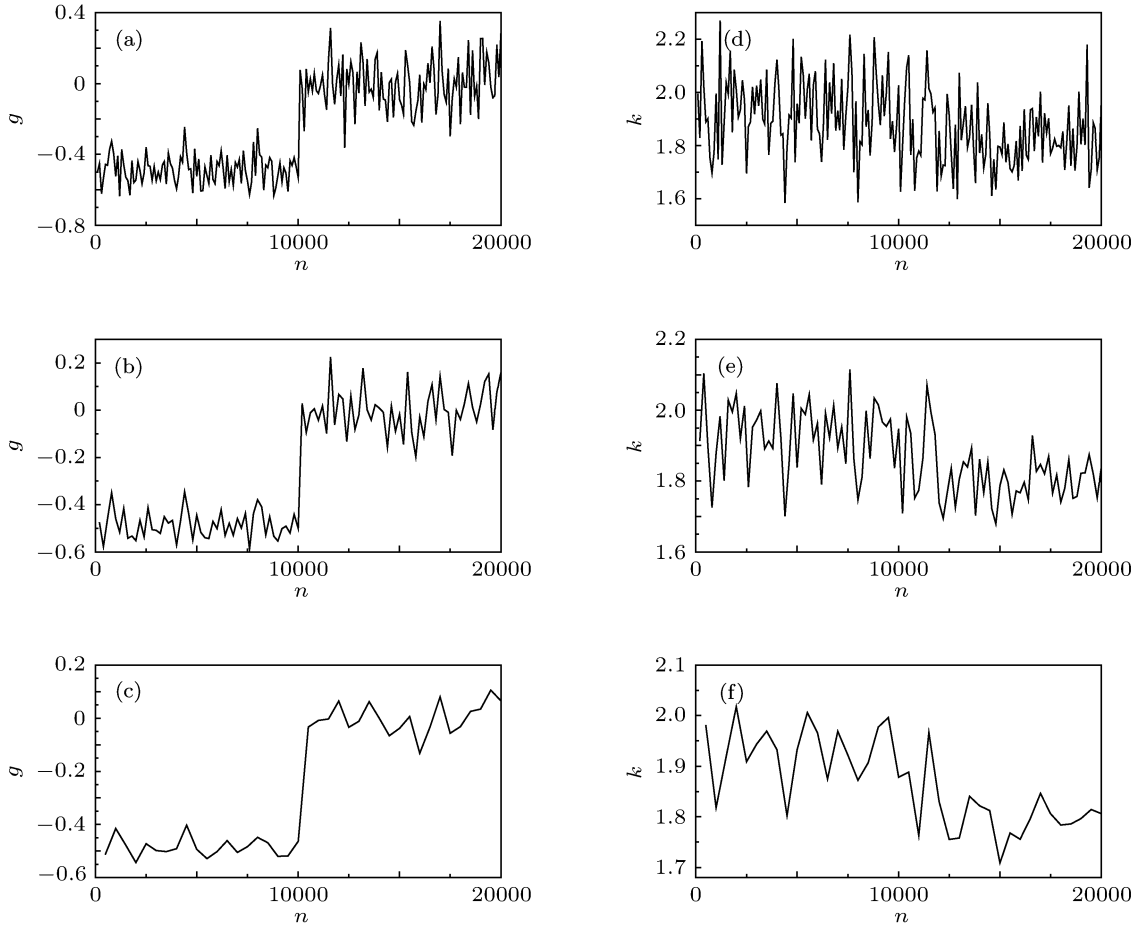

图 4 偏度系数和峰度系数对理想系列 IS2 的突变检测 (a), (b), (c) 分别为子序列长度为 100, 200, 500 时偏度系数的检测结果; (d), (e), (f) 分别为子序列长度为 100, 200, 500 时峰度系数的检测结果

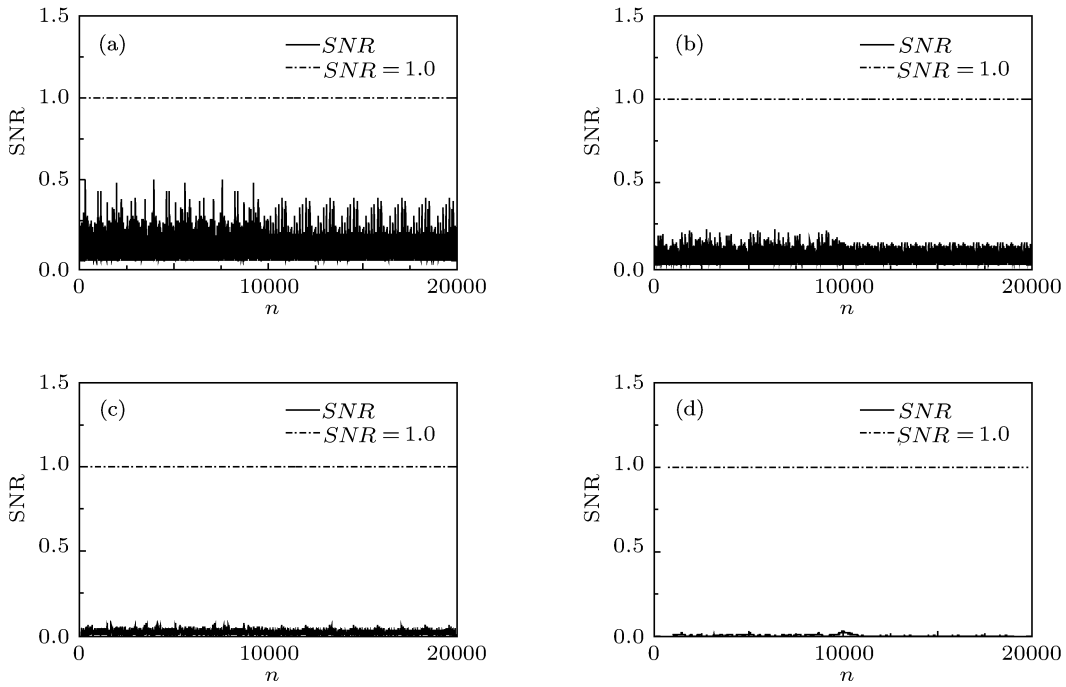

图 5 Yamamoto 信噪比法对理想时间序列 IS1 的突变检测结果, 当信噪比 (signal to noise ratio, SNR) 曲线超过 1.0 时, 则认为有突变发生 (a) 子序列长度为 10; (b) 子序列长度为 20; (c) 子序列长度为 100; (d) 子序列长度为 1000

可见, 偏度系数和峰度系数都能够准确地应用于检测理想序列 IS1 中存在的参数突变, 而且我们对其他类似的参数突变以及理想序列总长度较小 (如 2000, 3000, 5000 等) 时的情形进行了类似试验, 结果类似于 IS1, 不再赘述. 不同子序列长度下, 偏度系数和峰度系数检测结果的一致性, 表明了这两种方法的检测结果对于子序列长度的依赖性非常小, 有别于滑动 t- 检验, Crammer 法等传统突变检测方法的检测结果对于子序列长度的严重依赖性.

为了进一步测试偏度系数和峰度系数对于时间序列中突变信号的捕捉与识别能力, 本文考察了这两个系数在系统的动力学方程形式发生突变时的检测性能. 为此, 考虑如下一种情形: 在某个时刻, 由于一场突如其来的灾难, 种群的演变方程突然由 Logistic 方程 ((3) 式, 参数 u = 3.8) 突然变为随机行为 (由均匀分布的随机数模拟). 基于此, 本文构造了另一理想时间序列, 序列总长度为 20000, 该序列在 n = 10001 时方程形式发生了突变, 图 3(a) 展示了其演化曲线. 类似于 IS1, 分别对其在突变前后的数据进行了标准化处理, 即可得到新的理想序列 IS2 (图 3(b)).

从偏度系数对 IS2 的检测结果来看 (图 4(a)—(c)), 在不同子序列长度下, 在 n = 10001 前后, 偏度系数的演化展现出两种截然不同的稳定变化状态, 即在 n <10001 时, 偏度系数均明显小于 0, 表明了概率密度分布的左偏态势, 虽然随着子序列长度的增加, 偏度系数的波动幅度有所减小, 但是这种总的分布趋势没有发生任何变化. 而在 n >10000 后, 偏度系数突然变大, 且其值一直围绕着 0 附近波动变化, 根据偏度系数的物理意义可知, 这意味着概率分布密度曲线接近于正态分布时的情形. 因此, 可以判断在 n = 10001 前后, IS2 的概率密度分布型由左偏型突然转变为接近正态分布型. 但从 IS2 的偏度系数检测结果来看, 其值均小于 3, 这意味着 IS2 中概率密度分布曲线的陡峭程度较正态分布平坦. 从其子序列长度取 100, 200 和 500 时的检测结果来看, 在子序列长度较小时 (如子序列长度为 100), 峰度系数对于 IS2 中的动力学结构突变难以有效识别, 但是当子序列长度增加后 (如子序列长度取为 200 和 500), 峰度系数对于这种突变信号的识别还是能够有所反映. 进一步的研究表明, 当子序列长度足够大时 (如子序列长度取为 800 和 1000), 发现峰度系数对于 IS2 中的动力学结构突变还是能够有效识别 (图略). 这意味着, 就 IS2 的动力学突变检测而言, 偏度系数和峰度系数都能够进行有效识别, 只不过峰度系数对于子序列样本量的要求更高, 其原因主要在于 Logistic 模型种群的

概率分布曲线的陡峭程度与均匀分布随机数之间的这种差异性较小, 需要更多样本量才能够对此进行准确区分.

为了进一步证明新方法的有效性和优越性, 有必要对其与传统的突变检测方法进行比较. Yamamoto 信噪比方法 [8] 是比较严格的突变检测方法, 当信噪比大于 1.0 时, 其他方法也能检测出突变; 而当其他方法检测出有突变时 (甚至滑动 t- 检验达到 0.01 信度时), 信噪比很可能还达不到 1.0. M-K 方法 [6,7] 是比较不严格的, 因为它不考虑原始数据的数值, 只比较相临两个数值的大小, 所以在原理上有问题, M-K 方法经常会出现虚报突变的情况. 鉴于此, 本文首先将信噪比方法用于检测理想时间序列 IS1 中的突变, 以便同本文的新方法进行比较. 信噪比方法的检测结果表明, 无论子序列长度取较小值 (如 10, 20), 还是子序列长度取较大值 (如 100, 1000), 理想时间序列 IS1 和 IS2 中的信噪比水平均小于 1.0 (图 5), 这意味着信噪比方法未能够检测到 IS1 和 IS2 中存在的参数突变和动力学结构突变. Yamamoto 信噪比方法对 IS2 的检测结果类似于 IS1, 不在此赘述. 同时对其他传统突变检测方法如滑动 t- 检验法、Crammer 法、M-K 方法也进行了类似试验, 发现均未能对 IS1 和 IS2 中的突变信号进行有效检测 (图略). 对比偏度系数和峰度系数相应的检测结果可知, 新方法相对于 Yamamoto 信噪比方法等传统方法具有明显的优势.

4 结 论

本文基于一个稳定的动力系统其变量的概率密度具有较为稳定的分布型, 而当动力学结构发生变化后可能会导致系统变量的分布型发生不同程度的变化这一基本特征, 从系统变量的概率密度分布的微小变化角度出发, 将偏度系数和峰度系数用于捕捉和识别时间序列中的突变信号, 为突变检测提供了一条新的途径. 数值试验结果表明, 偏度系数针对参数突变和动力学结构突变都能够有效进行检测, 且其检测结果基本不依赖于子序列长度的选择, 而峰度系数也能够有效地识别时间序列中的参数突变和动力学突变. 而传统的突变检测方法如 Yamamoto 信噪比方法、滑动 t- 检验、Crammer 法、M-K 法等均未能有效识别出理想时间序列中的突变信息, 这意味着本文提出的突变检测新途径较 Yamamoto 信噪比方法等传统突变检测方法具有明显的优势. 同时也表明新方法在物理学、地学、气象学、社会科学等多学科中都有潜在的应用价值.

需要指出的是, 当突变前后概率密度曲线的偏斜性或陡峭程度相差较小时, 在运用偏度系数或峰度系数进行突变检测过程中, 可能需要更多的样本量, 即所选择的子序列长度需要更大一些, 以便确保偏度系数和峰度系数统计结果的可靠性, 从而能够有效地捕捉和识别时间序列中的突变信息. 因此, 新方法在某些情况下, 可能对于待检测时间序列的样本量的要求较高, 对于一些记录较少的时间序列可能并不适用, 这也是本文所提方法的缺陷之一. 在后续工作中, 将针对此缺陷展开进一步的深入研究.

作者衷心感谢匿名审稿专家的有益建议和意见, 感谢扬州大学物理科学与技术学院钱忠华老师和周云同学的有益讨论!

[1] Ye D Z, Tao S Y, Li M C 1958 *Acta Metro. Sin.* **29** 249 (in Chinese) [叶笃正, 陶诗言, 李麦春 1958 气象学报 **29** 249]

[2] Krishinamulti T N, Ramanathan Y 1982 *J. Atmos. Sci.* **39** 1290

[3] Marten S, Jordi B, William A B, Victor B, Stephen R C, Vasilis D, Hermann H, Egbert H V N, Max R, George S 2009 *Nature* **461** 53

[4] Li Y C, Yan W, Jiang C S, Zuo Y L 2009 *Recent Developments in World Seismology* **369** 1 (in Chinese) [李迎春, 闫伟, 蒋长胜, 左玲玉 2009 国际地震动态 **369** 1]

[5] He W P 2008 *Ph. D. Dissertation* (Lanzhou: Lanzhou University)(in Chinese) [何文平 2008 博士学位论文 (兰州: 兰州大学)]

[6] Mann H B 1945 *Econometrica* **13** 245

[7] Kendall M G, Charles G 1975 *Rank Correlation Methods* (New York: Oxford University) p202

[8] Yamamoto R, Iwashima T, Sanga N K 1985 *J. Meteor. Soc. Japan.* **63** 1157

[9] Livina V N, Havlin S, Bunde A 2005 *Phys. Rev. Lett.* **95** 208501

[10] He W P, Feng G L, Wu Q, Wan S Q, Chou J F 2008 *Non. Proc. Geophys*. **15** 601

[11] He W P, Wu Q, Zheng W, Cheng H Y 2010 *Acta Phys. Sin.* **59** 8264 (in Chinese) [何文平, 吴琼, 张文, 成海英 2010 物理学报 **59** 8264]

[12] Shi N, Gu J Q, Yi Y M, Lin Z N 2005 *Chin. Phys.* **14** 844

[13] Shi N, Yi Y M, Gu J Q, Xia D D 2006 *Chin. Phys.* **15** 2180

[14] Feng G L, Dong W J, Gong Z Q, Hou W, Wan S Q, Zhi R 2006 *Nonlinear Theories and Methods on Spatial-Temporal Distribu-*

tion of the Obserbational Data (Beijing: Metrological Press) pp27–89 (in Chinese) [封国林, 董文杰, 龚志强, 侯威, 万仕全, 支蓉 2006 观测数据非线性时空分布理论和方法 (北京: 气象出版社) 第 27— 第 89 页]

[15] He W P, Wu Q, Zheng W, Wang Q G, Zhang Y 2009 *Acta Phys. Sin.* **58** 2862 (in Chinese) [何文平, 吴琼, 张文, 王启光, 张勇 2009 物理学报 **58** 2862]

[16] Zhang W, Wan S Q 2008 *Chin. Phys.* B **17** 2311

[17] Wang Q G, Zhang Z P 2008 *Acta Phys. Sin*. **57** 1996 (in Chinese) [王启光, 张增平 2008 物理学报 **57** 1996]

[18] He W P, He T, Cheng H Y, Zhang W, Wu Q 2011 *Acta Phys. Sin.* **60** 813 (in Chinese) [何文平, 何涛, 成海英, 张文, 吴琼 2011 物理学报 **60** 813]

[19] Babak A S, Krishnaprasad P S 2004 *EURASIP J. Appl. Signal Processing* **15** 2295

[20] Abhisek U, Rastko Z 2006 *Electric Power Systems Research* **76** 815

[21] May R 1976 *Nature* **261** 459

A new approach to abrupt change detection based on change of probability density distribution*

Cheng Hai-Ying[1)] He Wen-Ping[2)†] He Tao[3)] Wu Qiong[4)]

1) (*Department of Fundamental Science Teaching, Yancheng Institute of Technology, Yancheng* 224051*, China*)

2) (*National Climate Center, China Meteorological Administration, Beijing* 100081*, China*)

3) (*Jinan Environmental Monitoring Center, Jinan* 250014*, China*)

4) (*National Satellite Meteorological Center, China Meteorological Administration, Beijing* 100081*, China*)

(Received 9 April 2011; revised manuscript received 18 May 2011)

Abstract

For a stable dynamic system, probability density distribution (PDD) of a system variable is relatively stable, and if there is a change in dynamic structure of a system, the PDD of the system variable will have some change correspondingly. According to this characteristic of PDD of a dynamic system, in this paper we present two new methods, namely, skewness index and kurtosis index, to detect an abrupt change in a time series by means of identifying some small changes in PDD. Tests on model time series indicate that skewness index and kurtosis index can be used to identify an abrupt change, such as abrupt change in parameter of an equation and abrupt dynamic change. Thus, we provide a new approach to detecting abrupt change in time series based on PDD. Further studies show that the detected results of the skewness index and kurtosis index are almost independence of the length of a subseries.

Keywords: coefficient of skewness, coefficient of kurtosis, probability density distribution, abrupt change detection

PACS: 92.60. Wc

* Project supported by the National Natural Science Foundation of China (Grant Nos. 40905034, 41175067, 40930952) and the Special Scientific Research Fund of Meteorological Public Welfare Profession of China (Grant Nos. GYHY201106015, GYHY201106016).

† E-mail: wenping_he @163.com