非线性时间序列的替代数据检验方法研究 [1]

雷 敏 王志中

(上海交通大学生物医学工程系 上海 200030)

摘 要 目前，替代数据方法已逐渐成为时间序列的非线性成分检验中广为采用的一种方法。但对于具有同原始数据的均值和方差的线性相关高斯过程的零假设，通常产生相应替代数据的 FT(Fourier Transform) 算法不能很好地重构原始数据的 Fourier 频谱。本文对替代数据方法进行了研究，提出了一种改进的 FT 算法，使得替代数据既具有原始数据的均值和方差，又具有原始数据的 Fourier 频谱。利用 Gauss 数据和 logistic 方程产生的混沌时间序列数据，证明本文提出的改进算法是可行的，所产生的替代数据是合适的。

关键词 时间序列，非线性成分，替代数据，零假设， Fourier 变换

中图号 TN911.72

1 引 言

目前在检验时间序列是否具有混沌动力学特性时，通常有两类方法：一类是直接识别时间序列数据中的混沌动力学特性，一类是通过检验数据中的非线性成分，间接地判断其混沌动力学特性。在直接方法中，已有很多原理上有效的方法，如计算关联维数、Lyapunov 指数和复杂度等，并已得到很多成功的应用，但这些方法的可靠性依赖于尽可能长的数据量，且易受测量噪声的干扰 [1] 。因此，对于有噪声的短数据，这些方法常会出现虚假的判断结果，尤其是对于低维混沌信号数据的检验。为了避免这些局限性， 1992 年， Theiler 等人提出了以替代数据 (surrogate data) 作为检验时间序列中非线性成分的方法 [2] ，即间接方法。该方法的基本思想是首先指定某种线性随机过程为零假设，并依据该假设产生相应的一组替代数据，然后分别计算比较原始数据和替代数据集的检验统计量，如果原始数据所算得值与替代数据集的值有显著差异，则拒绝该零假设，即该零假设不成立，说明原始数据中存在确定性的非线性成分。显然，间接方法比直接方法要容易得多。替代数据方法是目前检验时间序列非线性因素的重要经验方法，尽管只采用该方法还不能确定引起时间序列非线性的内在机制，但是当它与某些专门的算法比如混沌时间序列分析方法相结合使用时，就可以使两者的潜在能力得以充分发挥，为检验时间序列非线性的产生机理提供客观依据，因而替代数据方法自提出后就在有关混沌时间序列的研究中得到迅速而广泛的应用 [3-5] 和发展 [6-8] 。但作为产生替代数据的一种主要方法， FT(Fourier Transform) 算法目前还不能很好地重构原始数据的频谱特性 [2,8] ，文献 [2] 指出了这一算法存在的缺陷，这一缺陷的存在致使其应用有效性受到限制。为解决这一问题，本文对替代数据方法进行了研究，并提出了一种改进的 FT 算法，从而解决了目前 FT 算法所存在的问题，为替代数据的实际应用提供了一种新的算法。

2 替代数据方法

替代数据方法由零假设和检验统计量两部分组成。零假设是给出可以或不可以充分地说明数据本质的替代过程。本文首先回顾了现有的零假设及其产生替代数据的算法，针对目

[1] 1999-05-09 收到， 1999-11-15 定稿
国家自然科学基金资助项目 69675002

前广为采用的 FT 算法所存在的缺陷，给出了本文改进的 FT 算法，然后讨论了选择检验统计量的依据，并给出了本文所采用的检验统计量计算方法。

2.1 各种零假设及其算法 [2]

零假设 1　观测数据由相互独立而又分布相同的随机变量产生。

这里，通常采用高斯型的随机变量分布，所产生出的替代数据是时间独立的序列，且具有原始数据的均值、方差和幅值分布等特性，但是原来数据中任何时间关联性都已被破坏。相应的替代数据可以按下面方法得到：首先用伪随机数发生器形成高斯型白噪声，然后以噪声序列的秩或次序来重新排列实验数据，所得到的就是符合零假设 1 并与实验数据具有相同的均值、方差和幅值分布的替代数据。Schienkman 和 LeBaron 曾经将此方法用于股票市场的分析 [9] 。Breeden 和 Packard 也曾用此方法证明了时间上非一致采样的类星体时间序列数据中存在一定的动力学结构 [10] 。

零假设 2　观测数据由 Ornstein-Uhlenbeck 过程产生。

依据这一假设所产生的替代数据是具有最简单时间相关性的序列。Ornstein-Uhlenbeck 过程可以由以下方程经过迭代得到

$$x_t = a_0 + a_1 x_{t-1} + \sigma e_t \tag{1}$$

其中 e_t 是零均值、方差为 1 的高斯白噪声，系数 a_0 、 a_1 和 σ 一起决定时间序列 x_t 的均值、方差和自相关时间。其自相关函数为指数形式，令 $\lambda = -\log a_1$ ，用 $< \cdot >$ 表示对时间 t 取平均，则

$$A(\tau) \equiv \frac{< x_t x_{t-\tau} > - < x_t >^2}{< x_t^2 > - < x_t >^2} = e^{-\lambda|\tau|} \tag{2}$$

要由实验数据产生符合零假设 2 的替代数据，应该先计算原始数据的均值 μ 、方差 γ 和自相关函数 $A(1)$ ，再拟合方程的系数 $a_1 = A(1), a_0 = \mu(1 - a_1)$ 和 $\sigma^2 = \gamma(1 - a_1^2)$ ，然后通过上述方程迭代即可获得替代数据。

零假设 3　观测数据由具有原始数据的均值和方差的线性相关高斯过程产生。

这一零假设通常用于检验原始时间序列是否含有非线性成分，它可以用自回归 AR 模型表示：

$$x_t = a_0 + \sum_{k=1}^{q} a_k x_{t-k} + \sigma e_t \tag{3}$$

依据这一假设的替代数据，可由两种方法产生，一种方法是直接利用上式，通过方程的不断迭代而得到替代数据，但是用实验数据拟合 (3) 式中系数时会产生误差，致使迭代结果很容易发散，所以这种算法很不稳定；另一种方法是 Theiler 等人采用的将 Fourier 变换结果的相位进行随机化处理的方法，这种算法则比较稳定，它最先由 Osborne 等人提出。这种方法的思想是，通过重构原始数据的功率谱以保证替代数据同原始数据的线性相关性。文献 [2] 给出的步骤是先对实验数据进行 Fourier 变换，设观测数据为 $x(n)$ ，则它的离散 Fourier 变换为

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-2\pi i n k/N} \tag{4}$$

对所得到的各个频率处的变换值乘以 $e^{i\varphi}$ 进行相位随机化处理：

$$X'(k) = X(k) e^{i\varphi} \tag{5}$$

其中 φ 从区间 $[0,2\pi]$ 随机地选取，且满足斜对称条件 $\varphi(k)=-\varphi(N-k)$，使得 $X'(k)=\overline{X'(N-k)}$，以确保 Fourier 逆变换的结果是实数。再进行 Fourier 逆变换:

$$x'(n)=\frac{1}{N}\sum_{k=0}^{N-1}X'(k)e^{2\pi ink/N} \tag{6}$$

$x'(n)$ 即为所求的替代数据。但实际上这种算法的 Fourier 逆变换结果并非是实数，从图 1(a) 可以看出其 Fourier 逆变换的虚部数据比较大，是不能忽略不计的 (原始数据由 (10) 式产生)。这样所得的替代数据正如 Theiler 等人在文献 [2] 中所述的那样——不能较好地重构原始数据的 Fourier 频谱而且均值和方差也有一些差异。为此，Theiler 等人采用加窗 FT 算法对其进行了改进，随后，T. Schreiber 等人也对此作了进一步的研究 [8]，但仍没能很好地重构原始数据的 Fourier 频谱。

零假设 4 观测数据由线性相关的高斯噪声经静态非线性变换产生。

静态非线性变换是指观测或是测量函数具有非线性，静态是指 t 时刻观测的结果 x_t 只取决于该时刻动力过程的取值 y_t，而与以前时刻的值或者导数等无关。设观测函数为 h，则

$$x_t=h(y_t) \tag{7}$$

为了能产生替代数据，应进一步假设观测函数 h 是可逆的，这就使得这一零假设的应用受到一定的限制。其算法的详细步骤请参见文献 [2]。

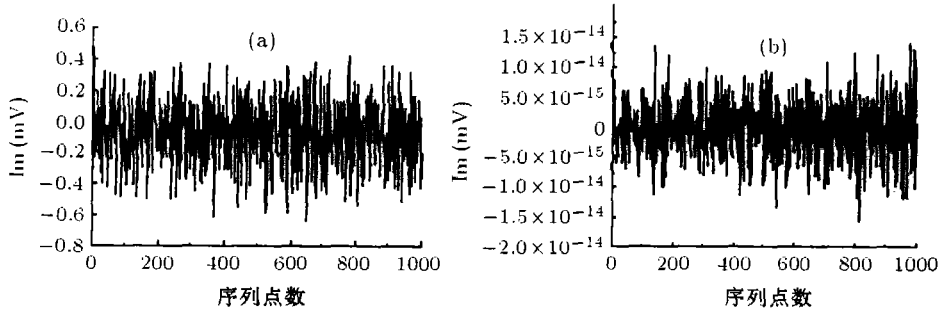

图 1
(a) 以前算法的 Fourier 逆变换后的虚部
(b) 本文算法的 Fourier 逆变换后的虚部

2.2 改进的 FT 算法

由上述的分析，根据零假设 3 产生替代数据的 FT 算法因其简单、稳定而广为采用，但其所存在的缺陷致使使用受到一定的限制 [11]。为此，本文提出一种改进的算法：在对原始数据进行 FT((1) 式) 后，随机化相位区间选为 $[-\pi,\pi]$，并且相位的斜对称条件将根据实际数据个数的奇偶性而有所变化。当数据为偶时，$\phi(f_0)=0,\ \phi(f_i)=-\phi(f_k),\ i=2\sim N/2,\ k=N\sim N/2+1,\ \phi(f_{N/2+1})=0$；为奇数时，$\phi(f_0)=0,\ \phi(f_i)=-\phi(f_k),\ i=2\sim(N+1)/2,\ k=N\sim(N+1)/2+1$，从而使原始数据相位随机化后仍满足实数 FT 的要求。这样再进行 Fourier 逆变换即得到所要的替代数据。从图 1(b) 可以看出这种方法的逆变换的结果是实数，其虚部数值很小 (10^{-14} 数量级)，可以忽略不计。

2.3 检验统计量 [7]

零假设是用于说明数据的不充分的潜在解释，而检验统计量是度量时间序列某些特性的数字量，它可以检验零假设是否成立。如果所观测数据的检验统计量与零假设所期望的值不同，则该零假设被拒绝，说明观测数据与零假设有本质不同。否则，该零假设不被拒绝，说明观测数据与零假设基本一致。检验统计量 T 可分为中枢性和非中枢性两种。一般，对于依据零假设的所有过程 F_ϕ，若检验统计量 T 的概率分布是一致的，则称 T 是中枢的，否则 T 是非中枢的。由于并非所有的检验统计量都有一致的效果，在选择检验统计量时最好选择中枢性的与替代数据产生方法无关的并且对于原始数据和替代数据是有差异的检验统计量；即对于一个所给定的假设，有任意 $F_i \in F_\phi$ 中的每个实现 z_i 和原始数据 z，使得 $T(z) \neq T(z_i)$。因此为了很好地比较原始时间序列与其替代数据间的差异，本文采用了下面两种检验统计量：(8) 式和 (9) 式。其中 (8) 式的 T 值不受均值和方差影响，这样可以判断除均值和方差外的一些线性结构特征是否一致；(9) 式可判断与均值和方差有关的关联性是否一致。

$$T = \overline{(x-\bar{x})^4} \Big/ \overline{((x-\bar{x})^2)^2} \tag{8}$$

$$T = \frac{1}{n-1}\sum_{i=1}^{n-1}(x(i)-\bar{x})(x(i+1)-\bar{x}) \Big/ \overline{(x-\bar{x})^2} \tag{9}$$

其中 $\bar{x}$ 为 x 的均值，n 为数据长度。

3 检验替代数据

为了证明由本文的 FT 算法产生的替代数据是合适的，并优于以前的 FT 算法，下面分别利用高斯数据和由 logistic 系统产生的混沌时间序列进行验证。

3.1 高斯数据检验

这里所选的高斯数据是由伪随机发生器产生的均值为 0、方差为 1 的随机数，然后依据零假设 3，由以前的 FT 算法和本文改进的 FT 算法计算出替代数据，各产生 1000 组。由图 2(a) 和 2(b) 可以看出原始高斯数据的检验统计 T 与替代数据 T 无显著差异 (T 由 (8) 式计算)，零假设不能被拒绝，说明两种算法所产生的替代数据在不考虑均值和方差时，与原始数据是等价的。

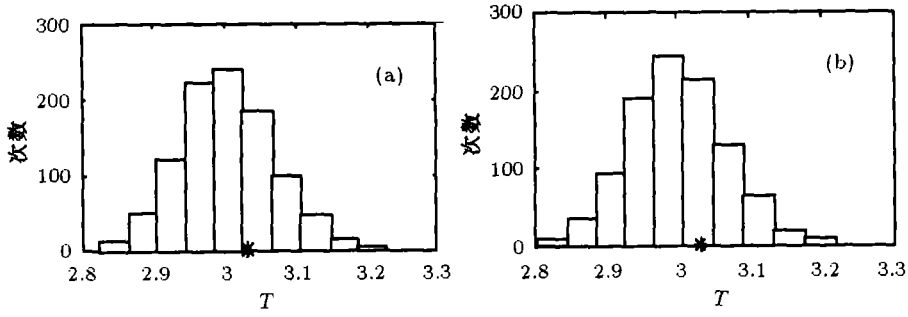

图 2　由 (8) 式计算的 T，直方图表示替代数据的 T 分布，* 表示原始数据的 T 值

(a) 替代数据由以前的 FT 算法产生

(b) 替代数据是由改进的 FT 算法产生

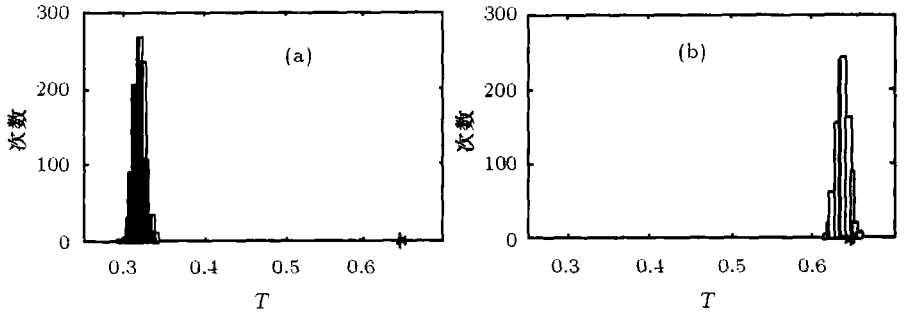

图 3 由 (9) 式计算的 T, 直方图表示替代数据的 T 分布, * 表示原始数据的 T 值

(a) 替代数据由以前的 FT 算法产生

(b) 替代数据是由改进的 FT 算法产生

但若考虑均值和方差时，情况则不同了，由 (9) 式所算得的 T 值分布就可以看出其差异，如图 3(a) 和 3(b) 所示，图 3(a) 的替代数据与原始序列是有差异的，说明以前的 FT 算法产生的替代数据不具有原数据的均值和方差，这是由于其不能很好地描述原始数据的线性特征——重构原始数据的 Fourier 频谱而造成的，这一点与算法原理分析一致，因此以前的 FT 算法所产生的替代数据不能很好地描述原始数据的线性结构；而本文改进的 FT 算法因能完全重构原始数据的 Fourier 频谱，故可很好地描述其线性结构，从而说明由本文改进的 FT 算法产生的替代数据优于以前的 FT 算法，该替代数据是与零假设完全一致的随机序列，即具有与原始数据相同的均值、方差和 Fourier 频谱。

3.2 混沌时间序列检验

为了进一步检验本文的 FT 算法所产生的替代数据适用于时间序列的非线性检验，本文运用混沌时间序列来检验。混沌是系统非线性达到一定程度时才出现的，所以混沌时间序列一定具有非线性。本文的混沌时间序列是由 (10) 式所描述的 logistic 系统当 $\alpha = 3.9$ 时产生的。

$$x_{t+1} = \alpha x_t(1 - x_t) + e_t \tag{10}$$

其中 $x_0 \in [0, 1]$, e_t 是白噪声，数据长度 N 为 5000 。替代数据由本文提出的改进 FT 算法产生，也是 1000 组。由图 4(a) 可以看出，混沌时间序列的检验统计量 T 与替代数据的 T 值都有显著不同 (由 (8) 式计算 T 值)，零假设 3 会被拒绝，说明除了均值、方差和 Fourier 频谱等线性特性一致外，替代数据和原始数据还有其他的特性不一致，从而表明混沌时间序列是非线性的，同时说明用本文改进的 FT 算法产生的替代数据能用于非线性检验，并说明对于非线性时间数据来说，只从频谱等线性性质上分析是不够的。另外，本文改进的 FT 算法计算的替代数据也能够用于短数据量的时间序列检验，如图 4(b) 为 $N = 500$ 时，替代数据和原始数据的检验统计量分布。显然，本文提出的方法为短数据时间序列的非线性分析提供了一种有效手段。

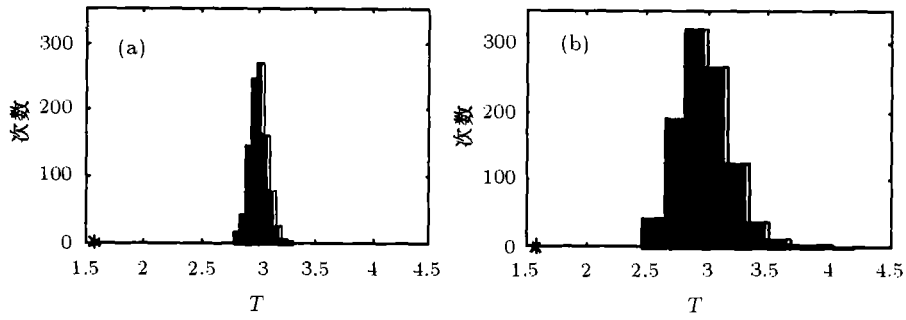

图 4　由 (8) 式计算的 T, 直方图表示替代数据, * 表示原始数据的 T 值

(a) 原始时间序列的长度 $N = 5000$

(b) 原始时间序列的长度 $N = 500$

4 结　论

通过对高斯数据和混沌时间序列的检验，说明本文提出的改进的 FT 算法是合理的，解决了文献 [2] 中所述 FT 算法的局限性，使得所产生的替代数据不仅具有原始数据的均值和方差，还能够完全重构原始数据的 Fourier 频谱。从而为产生具有与原始数据的均值和方差相同的线性相关高斯过程零假设的替代数据提供了可靠算法，且算法的稳定性很好，同时本文的替代数据方法也能用于短数据量的时间序列检验。

参　考　文　献

[1] K. Vibe, J. M. Vesin, On chaos detection methods, International Journal of Bifurcation and Chaos, 1996, 6(3), 529–543.

[2] J. Theiler, S. Eubank, A. Longtin, *et al.*, Testing for nonlinearity in time series, the method of surrogate data, Physica D, 1992, 58, 77–94.

[3] C. Poon, C. K. Merrill, Decrease of cardiac chaos in congestive heart failure, Nature, 1997, 389(10), 492–495.

[4] M. Barahona, C. Poon, Detection of nonlinear dynamics in short, noisy time series, Nature, 1996, 381(5), 215–217.

[5] U. Parlitz, L. Kocarev, Using surrogate data analysis for unmasking chaotic communication systems, International Journal of Bifurcation and Chaos, 1997, 7(2), 407–413.

[6] D. Prichard, J. Theiler, Generating surrogate data for time series with several simultaneously measured variables, Phys. Rev. Lett., 1994, 73(7), 951–954.

[7] J. Theiler, D. Prichard, Constrained-realization Monte-Carlo method for hypothesis testing, Physica D, 1996, 94, 221–235.

[8] T. Schreiber, A. Schmitz, Improved surrogate data for nonlinearity tests, Phys. Rev. Lett., 1996, 77(4), 635–638.

[9] J. A. Scheinkman, B. Lebaron, Nonlinear dynamics and stock returns, Journal of Business, 1989, 62(3), 311–337.

[10] J. L. Breeden, N. H. Packard, Nonlinear analysis of data sampled nonuniformly in time, Physica D, 1992, 58, 273–283.

[11] M. Small, K. Judd, Detecting nonlinearity in experimental data, International Journal of Bifurcation and Chaos, 1998, 8(6), 1231–1244.

STUDY OF THE SURROGATE DATA METHOD FOR NONLINEARITY OF TIME SERIES

Lei Min Wang Zhizhong

(*Dept. of Biomedical Engineering, Shanghai Jiaotong University, Shanghai* 200030, China)

Abstract Currently, the surrogate data method has become a widely used method in testing nonlinearity of time series. However, for the null hypothesis of linearly autocorrelated Gaussian noise with the mean and variance of the raw data, the exiting FT algorithms of generating surrogate data can not reproduce "pure" frequencies very well. In this paper, the surrogate data method is studied and an improved FT algorithms is proposed. Using the proposed algorithm, the surrogate data sets have the same mean, variance and Fourier spectrum with the original data. This FT algorithm is compared to the previous and proved feasible by using Guass series and chaos time series of the logistic system.

Key words Time series, Nonlinearity, Surrogate data, Null hypothesis, Fourier transform

雷 敏: 女，1968 年生，博士生，研究方向为医学信号处理、非线性时间序列分析等.
王志中: 男，1954 年生，教授，博士生导师，研究方向为医学信号处理、远程医疗等.