一种基于 KS 检验的时间序列非线性检验方法

侯澍旻 李友荣 刘光临
(武汉科技大学机械自动化学院 武汉 430081)

摘 要：检验统计量的选取将对时间序列非线性检验的结果产生重要影响。该文在采用打乱相位法产生替代数据后，引入了一种非参数检验——Kolmogorov-Smirnov 检验(简称 KS 检验)作为检验统计量。通过对各类信号的数值实验及与传统使用的高阶自相关量以及时间反演不可逆量对比结果表明，KS 检验是一种有效、稳定的非线性检验统计量，对噪声信号具有较强的抗噪能力，而对非线性信号具有较高的敏感性。

关键词：非线性检验；非线性时间序列；KS 检验；替代数据
中图分类号：TN911.72 **文献标识码**：A **文章编号**：1009-5896(2007)04-0808-03

A New Method of Detecting Nonlinear for Time Series Based on KS Test

Hou Shu-min Li You-rong Liu Guang-ling
(*College of Mechanical Engineering, Wuhan University of Science and Technology, Wuhan* 430081, *China*)

Abstract: The choice of test statistics can bring important influence to nonlinear of time series. This paper introduces a Non-parameter test——Kolmogorov-Smirnov (KS) test into nonlinearity test. After applying the Phase-randomized surrogate algorithm to create surrogate data, three test statistics methods, which include KS test, the third-order autocovariance and the asymmetry due to time reversal, are employed to determine nonlinear of five kinds signals. By comparing the test results, it indicates that KS test is an effective and stable nonlinear test statistics. The proposed method has high noise immunity and more sensitive to nonlinear signal.

Key words: Nonlinear test; Nonlinear time series; KS test; Surrogate data

1 引言

在自然界中存在大量形态复杂多样的时间序列，如太阳黑子序列、传感器拾取的设备信息序列、股票价格序列等。对于这些时间序列，为了判断其非线性，最直接的方法是通过计算关联维数、估计Lyapunov指数等混沌特征量来识别。其缺点是所需数据量大，计算结果受噪声影响大[1]。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份的方法[2]。

替代数据法是一种统计检验方法，它是根据随机系统中零假设的思想提出的。即假设测得的时间序列数据是线性的，并依据该假设产生相应的一组替代数据，然后分别计算原始数据和替代数据集的检验统计量。若两者有显著差异，则拒绝零假设，即原始时间序列不太可能是从与零假设相一致的系统中产生的，说明原始数据中应该存在确定的非线性成分[3]。因此，替代数据假设检验法由零假设，替代数据算法和检验统计量 3 部分组成[4, 5]。本文首先利用基于FT算法的替代数据方法生成替代数据，然后分别选取传统使用的阶自相关量(简称C3 法)，时间反演不可逆量(简称REV法)以及本文提到的Kolmogorov-Smirnov 检验(简称KS检验)对几种常见的时间序列仿真实例进行检验，结果发现KS检验对于各类信号都能正确判断，其稳定性较好。因此，KS检验是一种鲁棒性较好的非线性检验方法。

2 基于 FT 的替代数据法

2.1 零假设及其算法

零假设：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。若待测数据和替代数据的检验统计量之间差异显著，该零假设不成立，说明原始数据中必定包含非线性成分；反之，说明原始数据与替代数据一样，都是由线性随机过程产生的。

根据以上假设，目前常用的替代数据法有：打乱排列次序法、打乱相位法和高斯标度替代数据法。其中打乱排列次序法的主要思想是把原数据前后排列次序打乱，这样得到的替代数据与原数据相比具有相同的概率分布和统计属性；打乱相位法的思想是将原数据进行傅里叶变换得到功率谱，改变功率谱中各频率成份的相位，再进行逆傅里叶变换，这样得到的替代数据与原数据的功率谱相同，具有相同的线性相关性；而高斯标度替代数据法则对原序列相位加入高斯分布

2005-08-12 收到，2006-04-17 改回
湖北省自然科学基金(2005ABA287)，机械传动与制造工程湖北省重点实验室(2005A04)资助课题

的随机序列，再产生随机化相位高斯过程。

最常用的打乱相位法产生替代数据的步骤如下[6]：

(1) 输入测试时间序列 $s(t_i)$，$i=1,\cdots,N$，将其转换为复数的形式：

$$z(n)=s(n)+iy(n) \tag{1}$$

其中 $s(n)=s(t_i)$，$y(n)=0$，$n=1,\cdots,N$。

(2) 构建离散傅里叶变换

$$Z(m)=S(m)+iY(m)=\frac{1}{N}\sum_{n=1}^{N}z_n e^{-2\pi i(m-1)(n-1)/N} \tag{2}$$

(3) 构建一组随机相位的集合 $\phi_m\in[0,\pi]$，$m=2,3,\cdots,N/2$。

(4) 应用随机相位到离散傅里叶变换数据上

$$Z(m)'=\begin{cases}Z(m), & m=1,m=\dfrac{N}{2}+1\\ \left|Z(m)\right|e^{i\phi_m}, & m=2,3,\cdots,\dfrac{N}{2}\\ \left|Z(N-m+2)\right|e^{-i\phi_{N-m+2}}, & \\ & m=\dfrac{N}{2}+2,\dfrac{N}{2}+3,\cdots,N\end{cases}$$

(5) 求出 $Z(m)'$ 的逆傅里叶变换

$$z(n)'=s(n)'+iy(n)'=\frac{1}{N}\sum_{n=1}^{N}Z_m' e^{-2\pi i(m-1)(n-1)/N} \tag{3}$$

显然，经过这种变换得到的替代数据序列与原始数据序列具有相同的功率谱和自相关函数，而非线性相关性则被相位随机化过程去除。

2.2 KS 检验统计量简介

许多文献指出，检验统计量的选择将会直接影响判断结果的优劣。本文引入KS检验作为统计量[7]。

KS 检验基于经验分布是理论分布相容估计的原则。它用于描述两个独立统计样本的相似性。假设 $X_1,\cdots,X_m\overset{\text{iid}}{\sim}F(x)$，$Y_1,\cdots,Y_n\overset{\text{iid}}{\sim}G(x)$，且全样本独立，$F(x)$ 和 $G(x)$ 为连续分布函数，我们感兴趣的检验问题为：

$$H_0:F(x)\equiv G(x)\leftrightarrow H_1:F(x)\neq G(x) \tag{4}$$

由 Glivenko 定理知，用经验分布函数来近似理论分布函数是可行的，于是 Smirnov 用统计量：

$$D=\max_{i,j}\left\{\left|F_m(X_{(i)})-G_n(Y_{(j)})\right|\right\} \tag{5}$$

来检验上面的假设问题，其中 $F(x)$ 和 $G(x)$ 分别为 X 样本和 Y 样本的经验分布函数，$X_{(i)}$ 和 $Y_{(j)}$ 分别表示 X 样本和 Y 样本的顺序统计量，m,n 表示样本数；H_0 的拒绝域为其取最大值。统计量 D 所对应的显著性水平 p 由可靠性分布函数 Q_{ks} 表示：

$$\text{prob}(D)=Q_{ks}(\lambda)=2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2} \tag{6}$$

其中 $\lambda=\left[\sqrt{N_e}+0.12+\dfrac{0.11}{\sqrt{N_e}}\right]D$，$N_e=\dfrac{mn}{m+n}$。

显然，若两独立样本非常相似，统计量距离 $D\to 0$ 时，$p\to 1$，反之亦然。因此，KS 检验可以作为非线性检验的统计量。其检验思路为：经过打乱相位法产生的替代数据其非线性相关性已被去除。用该组数据和原始数据进行 KS 检验，取拒绝域为 0.05，若显著性水平 p>0.05,则认为两信号相同，说明原始数据具有线性特征；若显著性水平 p<0.05,则认为两信号不同，说明原始数据具有非线性特征。

3　仿真实验

采用该方法和文献[8]中介绍的两种常用传统时间序列非线性检验方法 C3 法以及 REV 法，进行对比计算检验。

$$\beta_{c3}=\frac{1}{N}\sum_{n=2\tau+1}^{N}(s_n\cdot s_{n-\tau}\cdot s_{n-2\tau}) \tag{7}$$

$$\beta_{\text{rev}}=\frac{1}{N}\sum_{n=\tau+1}^{N}(s_n-s_{n-\tau})^3 \tag{8}$$

式中 s_n 为所获取的时间序列，τ 为延迟时间。

采用 Lorenz 信号，ECG 信号，线性 AR 模型产生的信号，Henon 信号以及 sunspots 信号进行数值实验研究。每种信号采样点数都为 1000 点。用上节介绍的替代数据法分别生成 100 组替代数据。选取拒绝域为 0.05。检验结果如表 1 所示。表中 C3，REV 法为 100 组替代数据秩检验计算结果。其判别标准为：当计算的秩检验值介于 2.5 与 97.5 之间时，原始数据具有线性性；反之，则具有非线性性。表中 KS 检验的结果为 100 组替代数据分别与原始数据进行统计后检验值的累加值。其计算标准为：取拒绝域为 0.05 时，若拒绝原假设，即原始数据具有非线性性，此时检验值 q=1；反之，检验值 q=0(例如：表中的 100 表示 $\sum_{i=1}^{100}(q)_i=100$)。

对比 3 种检验方法知，KS 检验在对各类数据进行判别时，都能正确识别其特征；而 C3 法和 REV 法分别都出现了误判。

表 1　拒绝域为 0.05 时的检验结果

	C3	REV	KS
Lorenz	1	43	100
	非线性	线性	非线性
ECG	97	1	100
	线性	非线性	非线性
AR	65	67	0
	线性	线性	线性
Henon	100	1	100
	非线性	非线性	非线性
Sunspots	98	78	100
	非线性	线性	非线性

在选取上述信号迭加上信噪比 $c=0.5$ 的噪声：$\text{Noise}(c)=c\cdot\text{randn}$ (randn 为零均值，方差为 1 的高斯白噪声)。表 2 为检验结果。

表 2 各类信号迭加噪声后的检验结果

	C3	REV	KS
Lorenz	1	8	100
	非线性	线性	非线性
ECG	92	74	92
	线性	线性	非线性
AR	42	53	0
	线性	线性	线性
Henon	100	1	100
	非线性	非线性	非线性
Sunspots	100	77	100
	非线性	线性	非线性

显然，KS 检验对加入噪声后的各类信号均能有效地检验出其线性特性，但采用 C3 法和 REV 法都会出现不同程度的误判。

4 结束语

检验统计量的选取对替代数据法的检验结果影响较大。本文尝试将 KS 检验引入非线性检验中，通过对弱非线性信号、强非线性信号以及加入噪声后信号的检验，该方法均能得到正确的判断结果。与传统方法对比结果表明，KS 检验是一种有效、稳定的非线性检验方法，对噪声信号具有较强的抗噪能力，对非线性信号具有较高的敏感性。

参 考 文 献

[1] Kanty H and Schreiber T. Nonlinear Time Series Analysis[M]. Cambridge: Cambridge University Press, 1997: 92–104.

[2] Theiler J, Eubank S, and Longtin A. Testing for nonlinearity in time series: The method of surrogate data [J]. *Physics D*, 1992, 58: 77–94.

[3] 孙海云, 王峰. 检验统计量的选取对替代数据方法的影响[J]. 数据采集与处理, 2003, 18(3): 253–257.
Sun Hai-yun and Wang Feng. Effect on choice of test statistics for surrogate data tests. *Journal of Data Acquisition & Processing*, 2003, 18(3): 253–257.

[4] 雷敏, 孟光, 冯正进. 连续动力系统时间序列的非线性检验[J]. 物理学报, 2005, 540(3): 1056–1063.
Lie Min, Meng Guang, and Feng Zheng-jin. Detecting the nonlinearity for time series sampled from continuous dynamic systems. *Acta Physica Sinica*, 2005, 540(3): 1056–1063.

[5] 雷敏, 王志中. 非线性时间序列的替代数据检验方法研究[J]. 电子与信息学报, 2001, 23(3): 248–254.
Lei Min and Wang Zhi-zhong. Study of the surrogate data method for nonlinearity of time series. *Journal of Electronics & Information Technology*, 2001, 23(3): 248–254.

[6] Schreiber T and Schmitz A. Surrogate time series[J]. *Phys. D*, 2000, 142: 346–382.

[7] 吴善元, 王兆军. 非参数统计方法[M]. 北京: 高等教育出版社, 1996: 144–150.

[8] Gautama T, Mandic D P, and Van Hulle M M. The delay vector variance method for determinism and nonlinearity in time series[J]. *Phys. D*, 2004, 190: 167–176.

侯澍旻： 男，1974 年生，讲师，博士生，研究方向为非线性时间序列分析及数据挖掘算法.

李友荣： 男，1946 年生，教授，博士生导师，研究方向为时间序列分析及智能故障诊断.

刘光临： 男，1946 年生，教授，博士生导师，研究方向为时间序列分析及智能故障诊断.