第 36 卷第 1 期 数学的实践与认识 Vol. 36 No. 1
2006 年 1 月 MATHEMATICS IN PRACTICE AND THEORY Jan., 2006

数据预处理对替代数据检验方法的影响

孙海云[1], 王 峰[2]

(1. 上海应用技术学院数理系, 上海 200235)
(2. 上海交通大学图像通信与信息处理研究所, 上海 200030)

摘要: 替代数据检验法是检验时间序列中是否存在确定性非线性成分的重要统计方法. 通过研究差分和数据平滑运算对替代数据检验方法的影响, 指出常用的线性滤波等数据预处理步骤破坏了序列的静态性质, 从而会导致对零假设的错误拒绝. 因此, 建议应该直接利用原始时间序列而非应用了差分等非静态滤波运算后的时间序列生成替代数据, 再进行假设检验, 以免造成对零假设的错误拒绝.

关键词: 替代数据; 非线性时间序列; 零假设; 差分; 滤波

1 引 言

替代数据检验方法是检验时间序列中是否存在非线性成分的重要统计方法. 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, 然后分别计算原始数据和替代数据集的检验统计量, 若二者有显著差异, 则拒绝零假设, 即原始时序不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性的非线性成分[1].

实测信号与观测数据的平滑与滤波等数据预处理步骤是观测数据分析中的一项重要准备工作. 如线性差分就常作为时序分析的准备步骤用来去除时序中的非平稳项及原始数据之间的线性相关性, 这种线性关联会在计算关联维数和其它非线性特征量时导致错误结果[2]. 但在替代数据检验方法中, 我们发现常用的线性差分及数据平滑等预处理步骤由于破坏了序列的静态性质而会导致对零假设的错误拒绝. 本文将结合常用的两种零假设及相应的替代数据算法, 来研究线性差分等数据预处理步骤对替代数据检验方法的影响.

2 替代数据假设检验方法

替代数据假设检验方法主要由零假设、替代数据生成算法和检验统计量等几部分构成.

2.1 常用零假设

零假设 1: 原始数据由线性高斯随机过程所产生.

该零假设可用来检验原始时序中是否含有非线性成分, 它可用自回归过程AR(P)模型表示:

$$x_{t+1} = \mu + \sum_{j=0}^{p-1} a_j x_{t-j} + \sigma e_t \tag{1}$$

其中, μ 和 σ^2 为原始数据的均值与方差, e_t 为均值为零、方差为 1 的高斯白噪声.

零假设 2: 原始数据为线性高斯随机过程AR(P)经过单调静态非线性变换 $g(\cdot)$ 所生

收稿日期: 2003-10-18

成, 即

$$\{\tilde{x}(t)\} \sim AR(P), \quad x(t) = g(\tilde{x}(t)) \tag{2}$$

在替代数据假设检验方法中, 零假设 2 尤为重要. 因为我们从实验或自然界中获得的原始时间序列很少有真正服从高斯分布的, 但我们却不能由此判断该序列存在非线性, 因为很多线性高斯时序会通过非线性观测函数来获得, 即其本质为线性的. 对零假设 2 的检验, 会帮助我们判断原始时序中的非线性成分是源于测量函数还是其动力学过程本身.

2.2　替代数据生成算法

由零假设 1 生成替代数据的方法为: 对给定的原始时间序列应用离散 Fourier 变换, 得

$$X(f) = F\{x(t)\} = \sum_{n=0}^{N-1} x(t_n) e^{2\pi i f n \Delta t} = A(f) e^{i\Phi(f)} \tag{3}$$

其中, $A(f)$ 和 $\Phi(f)$ 分别表示振幅和相位. 对所得到的变换值进行相位随机化处理, 将 $\Phi(f)$ 随机的旋转一相位角 $\varphi(f)$, 得到

$$\tilde{X}(f) = A(f) e^{i[\Phi(f) + \varphi(f)]} \tag{4}$$

其中 $\varphi(f) \in [0, 2\pi)$, 且满足条件 $\varphi(f) = -\varphi(-f)$, 以保证 Fourier 逆变换的结果为实数. 然后经 Fourier 逆变换即可得到替代数据

$$\tilde{x}(t) = F^{-1}\{\tilde{X}(f)\} = F^{-1}\{X(f) e^{i\varphi(f)}\} \tag{5}$$

不难看出, 经过这种变换得到的替代数据序列 $\{\tilde{x}(t)\}$ 与原始数据序列 $\{x(t)\}$ 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掉了[3]. 图 1(a) 为根据 AR(1) 序列生成的一组替代数据. 若对零假设 2 中非高斯时序直接采用上述生成替代数据, 则会因为原始时序与替代时序所服从的分布不同而导致对零假设的错误拒绝. 为此应对算法做适当调整[4], 以保证替代数据序列与原始序列的幅值分布相一致.

2.3　替代数据的非线性检验

利用替代数据生成算法, 生成 m 个替代数据集, 对原始数据和替代数据集同时采用非线性分析方法, 得到原始数据的非线性统计量值 q_0, 替代数据集的非线性统计量值 $q_1, \cdots, q_m$, 若 q_0 与 $q_1, \cdots, q_m$ 有显著不同, 则拒绝零假设.

时间不可逆性是非线性的重要特征, 因此本文采用时间反演不可逆量 q^{τ}_{REV} [5]

$$q^{\tau}_{REV} = \left\{ \left(\frac{x_i - x_{i+\tau}}{x_i - x_{i+\tau}} \right)^3 \right\}_2 \tag{6}$$

来检验时序的非线性, 其中 τ 为延迟时间, $\langle \cdot \rangle$ 表示均值, 它刻画了时间反演时时序的不对称性.

3　数据预处理对替代数据检验方法的影响

在时间序列分析方法中, 对原始时序进行线性滤波是一种常见的数据预处理方法, 可用来消除资料中的噪声, 但在替代数据检验方法中, 我们发现平滑、差分及维纳滤波等数据降噪方法常会导致对零假设的错误拒绝. 为此, 我们利用最简单的差分运算来研究数据预处理对替代数据检验方法的影响.

我们采用简单的一阶自回归 AR(1) 模型:

$$x_n = 0.99 \cdot x_{n-1} + \eta_n$$

其中, η_n 是均值为 0, 方差为 1 的高斯白噪声. 给定单调静态非线性函数 S:

$$y_n = (x_n)^3$$

则根据前面所介绍的替代数据方法, 序列$\{x_n\}$ 和$\{y_n\}$ 应分别接受零假设 1 和零假设 2 分别根据$\{x_n\}$ 和$\{y_n\}$ 生成 100 组替代数据, 检验结果与我们所设想的一致, 均接受原假设(见图 1(b)、图 1(d)).

若先对序列$\{x_n\}$ 做一次线性差分运算$x_t' = x_t - x_{t-1}$生成序列$\{x_n'\}$, 然后对$\{x_n'\}$ 生成 100 组替代数据进行检验, 结果仍是接受零假设 1(见图 1(c)), 即线性差分不会影响对零假设 1 的检验, 其原因为差分是对线性序列作线性运算, 因此不会改变序列的线性性质

但若对序列$\{y_n\}$ 做一次线性差分运算$y_t' = y_t - y_{t-1}$生成序列$\{y_n'\}$, 然后对$\{y_n'\}$ 生成 100 组替代数据进行检验, 结果却是拒绝零假设 2(见图 1(e)), 即认为时序的非线性性质是其本身内在非线性的表现, 而不是由线性序列经非线性变换造成的, 这与我们已知的序列$\{y_n\}$ 的性质是不一致的

先对$\{x_n\}$ 进行差分运算生成序列$\{x_n'\}$, 然后对差分序列$\{x_n'\}$ 进行单调静态非线性变换生成序列

$$y_n'' = (x_n')^3$$

对$\{y_n''\}$ 生成 100 组替代数据, 检验结果为接受零假设 2(见图 1(f)), 既可以正确的反映出原序列$\{x_n\}$ 的确是线性高斯序列经单调静态的非线性变换产生的, 而并非是系统本身内在非线性的表现

造成该错误的原因应该为零假设 2 中要求对线性序列做单调静态的非线性变换, 而差分运算会改变非线性变换的静态性质 为验证的确是该原因造成了对零假设 2 的错误拒绝, 我们再设计一种检验方式: 在数据处理中, 常利用数据平滑方法提取或消除数据中的趋势项, 这是对数据进行预处理的重要方法 在具体应用中, 不同的拟和观测点数和多项式次数, 会产生不同的光滑效果 本文采用简单的三点滑动平均公式

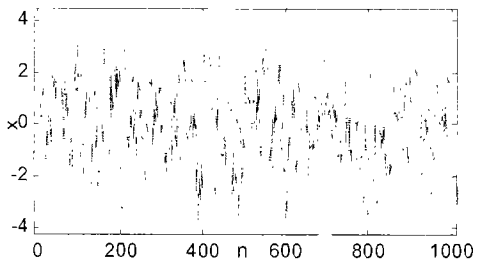
$$x_i = \frac{1}{3}[x_{i-1} + x_i + x_{i+1}]$$

来考察数据平滑处理对替代数据方法的影响 生成两种序列: 一是对进行了上述平滑处理后的一阶自回归AR (1)序列做三次方运算, 生成序列$\{y_n'''\}$; 一是对AR (1)模型的三次方序列进行平滑处理, 记为$\{x_n'''\}$. 从图 1(g)可看出, 利用零假设 2 对序列$\{y_n'''\}$ 的检验是正确的, 即$\{y_n'''\}$ 应为线性序列经非线性变换产生的时间序列 而对序列$\{x_n'''\}$ 的检验结果是拒绝零假设 2(见图 1(h)), 这显然是错误的

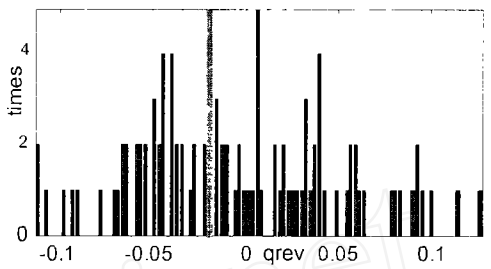
由于在实际应用中, 我们所获得的或测量到的往往只是经非线性变换后的序列$\{y_n\}$, 而不清楚变换前的确切形式, 所以无法直接利用最后一种检验方式进行检验, 即对需要检验的序列先差分后变换 因此, 我们建议应该直接利用获得的原始时间序列生成替代数据进行假设检验

4 结 论

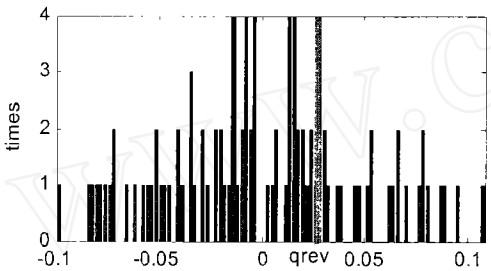
替代数据假设检验方法可用来判断原始数据中是否存在确定性的非线性成分, 从而为进一步的时间序列分析和建模工作提供方向性的指导. 但在替代数据检验方法中, 我们发现常用的线性差分及数据平滑等预处理步骤由于破坏了序列的静态性质而会导致对零假设的错误拒绝 因此, 本文建议应该直接利用原始时间序列而非应用了差分等非静态滤波运算后的时间序列生成替代数据, 再进行假设检验, 以免造成对零假设的错误拒绝



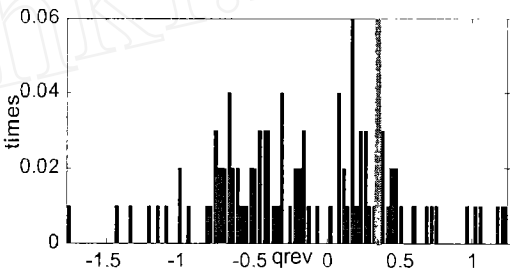
（a）AR(1)序列的一组替代数据



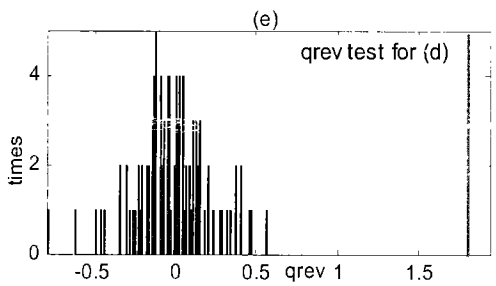
（b）序列 x_n 的检验结果



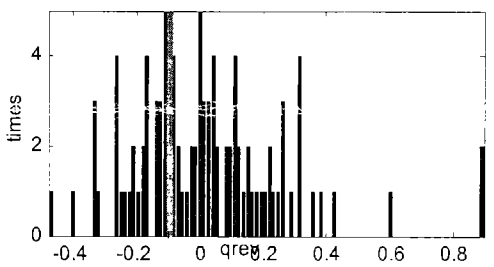
（c）序列 x_n' 的检验结果



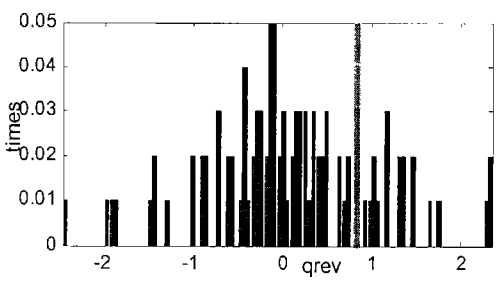
（d）序列 y_n 的检验结果



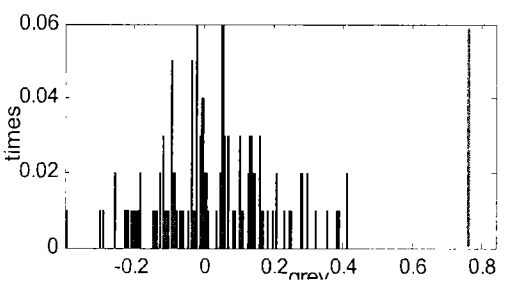
（e）序列 y_n' 的检验结果



（f）序列 y_n'' 的检验结果



（g）序列 y_n''' 的检验结果



（h）序列 x_n''' 的检验结果

图 1　统计量 q_{REV} 以各序列的检验结果

参考文献:

[1] Theiler J, Eubank S, Longtin A, Galdrikian B, Farmer J D. Testing for nonlinearity in time series: the method of surrogate data[J]. Physica D, 1992, 58: 77—94.

[2] Prichard D. The correction dimension of differenced data[J]. Phys lett A, 1994, 191: 245—250.

[3] Schreiber T. Interdisciplinary application of nonlinear time series methods[J]. Physics Report, 1998, 308(1): 51—59.

[4] Schreiber T, Schmitz A. Surrogate time series[J]. Physica D, 2000, 142: 346—382.

[5] Timmer J. The power of surrogate data testing with respect to nonstationarity[J]. Phys Rev E, 1998, 58: 5153—5156.

Effect of the Data Processing for Surrogate Data Tests

SUN Hai-yun[1], WANG Feng[2]

(1. Dept. of Mathematics and Physics, Shanghai Institute of Technology, Shanghai 200235, China)

(2. Institute of Image Communication and Information Processing, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: In the analysis of time series, the surrogate data test is often performed in order to investigate nonlinearity in the data. When we use this method, we find that the test does not always succeed to reject the null hypotheses when we apply the differencing or smoothing operator to the original data sets. Because it is not a static nonlinear transform of a linear process. So it is suggested that one should creating the surrogate data sets directly from the original time series, otherwise spurious results can occur.

Keywords: surrogate data; nonlinear time series; null hypothesis; difference; filter