

噪声为单位根过程的非参数函数变点的小波检测

齐培艳^{1,3}, 田 铮^{1,2}, 段西发^{1,3}, 李晓燕^{1,2}, 陈占涛^{1,2}

(1. 西北工业大学 应用数学系, 陕西 西安 710072;

2. 中国科学院 遥感应用研究所 遥感科学国家重点实验室, 北京 100101;

3. 太原科技大学 数学系, 山西 太原 030024)

摘要: 利用小波方法和极限定理对噪声为单位根过程的非参数函数的跳跃点进行检验. 首先, 利用极限定理得到噪声的小波系数的极限分布. 然后构造检验统计量, 在原假设成立的条件下得到任意尺度上检验的临界值, 证明了检验的一致性, 并给出小波系数的阈值; 在备择假设成立的条件下, 给出变点个数、变点位置的相合估计与收敛速度. 最后利用模拟研究与实例分析说明了方法的有效性和实用性, 并与“UNI”方法以及“GOF”方法作比较, 说明对于噪声为单位根过程的非参数函数变点的检测问题, 本文所提的方法更加有效.

关键词: 变点; 单位根过程; 小波变换; 收敛速度

中图分类号: O23, O29

文献标识码: A

Wavelet detection of jumping points in a nonparametric function with the unit-root noise

QI Pei-yan^{1,3}, TIAN Zheng^{1,2}, DUAN Xi-fa^{1,3}, LI Xiao-yan^{1,2}, CHEN Zhan-tao^{1,2}

(1. Northwestern Polytechnical University, Department of Applied Mathematic, Xi'an Shaanxi 710072, China;

2. National Key Laboratory of Remote Sensing,

Institute of Remote Sensing Applications Chinese Academy of Science, Beijing 100101, China;

3. Taiyuan University of Science and Technology, Department of Mathematics, Taiyuan Shanxi 030024, China)

Abstract: We combine the wavelet method and the limit-theorem to test the presence of jumping-points in a nonparametric function which is observed with unit-root noise. First, the limit distribution of the wavelet coefficients of the noise is derived using the limit-theorem, and then, the statistic of detection is determined. When the null hypothesis holds, we obtain the critical values at any scale, prove the consistency of wavelet detection and give the threshold of wavelet coefficients. When the alternative hypothesis holds, the consistent estimation of the numbers and locations of jumping points are given and the rate of convergence is obtained. Simulation study and real data analysis support our method. Finally, we compare our method with the “UNI”-method and the “GOF”-method. Asymptotic results show that our method is more powerful in detecting the jumping points of nonparametric function with Unit-root noise.

Key words: jumping-point; unit-root process; wavelet transformation; the rate of convergence

1 引言(Introduction)

变点问题起源于质量控制, 从生产线上抽检产品以检测产品质量是否超过其质量控制警界线时, 质量超过警界线的时刻称为变点. 变点问题在工业自动化控制、金融、医学等领域都有着实际应用背景^[1,2]. 对工序质量的检查、故障检测、分析汇率的变化规律、心电图的韵律分析、试飞数据中数据跳点的估计等都可以归结为对系统的变点研究.

变点问题的研究方法主要有极大似然方法^[3]、最

小二乘方法^[4]、累积和方法^[5]等. 小波变换具有自动改变窗长的功能, 可以很好地把信号在空间和频率上局部化, 这一特性使其成为变点分析中有力的数学工具. Wang(1995)^[6]基于小波系数的绝对值对含噪声信号的变点进行检验; Odgen和Parzen(1996)^[7]利用小波系数的累积和对变点进行检验, Raimondo 和Tajvidi(2004)^[8]应用广义Pareto分布对小波系数进行建模, 解决了方差无穷的厚尾信号的变点检验问题.

变点分析本质上属于数理统计和非线性时间序列分析范畴,它引起了国内外众多学者的关注^[3,9].非线性时间序列可分为平稳时间序列和非平稳时间序列两类,实际问题中很多数据都属于非平稳时间序列,如货币的供应量、物价指数等^[10].因此,对非平稳时间序列的变点研究有很重要的实际意义.值得注意的是单位根过程是常见的非平稳过程,且方差无穷,在金融系统工程中有广泛的应用.至今鲜见对噪声为非平稳时间序列的非参数函数的变点问题进行研究的文献.

2 模型与假设(Model and assumption)

考虑如下模型:

$$Y_i = f(i/T) + \epsilon_i, \quad i = 1, \dots, T,$$

其中: $\epsilon_i (i = 1, \dots, T)$ 是单位根序列, $\epsilon_i = \epsilon_{i-1} + u_i$; 函数 f 及 u_i 分别满足以下假设:

假设 1 $f(x)$ 为未知非参数函数, 变点个数、位置及跳跃度都是未知的, 而变点个数的上限 m 是已知的.

假设 2 $u_i = \varphi(L)e_i = \sum_{j=1}^{\infty} \varphi_j e_{i-j}$ 是平稳过程,

其中 $\{\varphi_j\}$ 满足 $\sum_{j=0}^{\infty} j|\varphi_j| < \infty$, $\{e_i, i = 1, 2, \dots\}$ 是独立同分布随机变量序列, 满足 $E(e_i) = 0$, $D(e_i) = \sigma^2 < \infty$.

本文研究如下假设检验问题:

H_0 : f 是光滑函数(在区间 $[0, 1]$ 上至少连续可微);

$H_1(m)$: f 至少有 1 个最多有 m 个跳跃点, 除此而外, f 是光滑的.

3 小波变换及噪声小波系数的极限分布(Wavelet transformation and the limit distribution of wavelet coefficients of noise)

令 $\psi(u)$ 是一个小波母函数, 对离散数据集 $Y = (Y_1, \dots, Y_T)$ 进行离散小波变换得到 Y 的经验小波系数

$$w'_{j,k} = w'_{j,k}(f) + w'_{j,k}(\epsilon),$$

其中:

$$w'_{j,k}(f) = \frac{1}{\sqrt{T}} \sum_{i=1}^T \psi_{j,k}(i/T) f(i/T), \quad (1)$$

$$w'_{j,k}(\epsilon) = \frac{1}{\sqrt{T}} \sum_{i=1}^T \psi_{j,k}(i/T) \epsilon_i. \quad (2)$$

参见文献[11].

对函数 f 进行小波变换得到的经验小波系数有如下基本性质^[12]:

1) 在原假设 H_0 下, f 是可微的, 对所有的分辨率水

平 $j \geq j_0 = 3$ 以及所有的 $k = 0, 1, \dots, 2^j - 1$, 有 $|w'_{j,k}(f)| \leq c_1(T2^{-3j})^{\frac{1}{2}}$.

2) 在备择假设 H_1 下, 若存在一个点 $x \in [k/2^j, (k+1)/2^j]$ 并且 x 是 f 的跳跃点, 则有 $|w'_{j,k}(f)| \geq c_2(T2^{-j})^{\frac{1}{2}}$.

常数 c_1, c_2 仅仅依赖于函数 f , 不会影响检验统计量的设定及检验结果, 为叙述方便, 本文假设 $c_1 = c_2 = 1$. 以上两条性质表明, 使得小波系数 $w'_{j,k}(f)$ 的绝对值达到最大的点正好对应着函数 f 的变点.

本节用 Haar 小波对数据进行小波变换, 并利用极限定理推导出噪声小波系数的极限分布.

Haar 小波定义为

$$\psi(x) = I_{[0, 1/2)}(x) - I_{[1/2, 1]}(x),$$

其基函数 $\psi_{j,k}(x)$ 的支撑是二进区间 $I_k = [k/2^j, (k+1)/2^j]$, 且在此区间中 i/T 的个数为 $T/2^j$, 因此有

$$w'_{j,k}(\epsilon) = \frac{1}{\sqrt{T}} \sum_{i \in I_k} 2^{j/2} (\pm) \epsilon_i = \frac{1}{\sqrt{T/2^j}} \sum_{i \in I_k} \epsilon_i, \quad (3)$$

应注意 ϵ_i 服从单位根过程, 方差无穷, 故利用以下结果对噪声的小波系数进行建模(参见文献[10]).

引理 1 设平稳过程 $\{u_i\}$ 满足假设 2. 令 $\xi_i = u_1 + u_2 + \dots + u_i, i = 1, 2, \dots, T$.

若 $\xi_0 = 0$, 那么

$$T^{-3/2} \sum_{i=1}^T \xi_{i-1} \Rightarrow \sigma \varphi(1) \int_0^1 W(r) dr \sim N(0, \frac{1}{3} \varphi(1) \sigma^2),$$

其中 $\varphi(1) = \sum_{j=0}^{\infty} \varphi_j$.

为叙述方便, 在给定的多分辨率水平 j 上, 令 $w_{j,k} = \frac{1}{T/2^j} w'_{j,k}$, 即

$$w_{j,k}(f) = \frac{1}{T^{3/2}/2^j} \sum_{i=1}^T \psi_{j,k}(i/T) f(i/T), \quad (4)$$

$$w_{j,k}(\epsilon) = \frac{1}{T^{3/2}/2^j} \sum_{i=1}^T \psi_{j,k}(i/T) \epsilon_i. \quad (5)$$

由引理 1 可得, 当 $T/2^j \rightarrow \infty$ 时,

$$w_{j,k}(\epsilon) \Rightarrow \sigma \varphi(1) \int_0^1 W(r) dr \sim N(0, \frac{1}{3} \varphi(1) \sigma^2). \quad (6)$$

4 主要结果(Main results)

4.1 变点检测(Detection of change-points)

本节旨在构造检验统计量并确定小波系数的阈值. 由上节中经验小波系数的基本性质可得如下定理:

定理 1 在水平 j 上, 对 $|w_{j,k}|$ 进行排序得到 $|w_{j,(1)}| > |w_{j,(2)}| > \cdots > |w_{j,(2^j)}|$, 令 $v_T = 2^{-j/2}T^{-1/2}$, 在原假设 H_0 成立时, 若多分辨水平 $j = j(T)$ 满足

$$2^j \asymp \left(\frac{T}{\log T}\right)^\delta, \quad 1 < \delta < 2, \quad (7)$$

则对任意的常数 $m \geq 1$, 当 $T \rightarrow \infty$ 时, 有

$$P(|w_{j,(m)}| > v_T) \rightarrow 1. \quad (8)$$

关系式 $a_T \asymp b_T$ 表示, 对于充分大的 T , 存在正常数 d_1 和 d_2 , 使得 $d_1 b_T \leq a_T \leq d_2 b_T$ 成立.

考虑如下检验统计量:

$$N_i = |w_{j,(i)}| - |w_{j,(m+1)}|, \quad i = 1, 2, \dots, m. \quad (9)$$

在原假设 H_0 成立时, 由定理 1 可得, $N_i < |w_{j,(i)}| - v_T, i = 1, 2, \dots, m$ 是依概率 1 成立的, 因此定理 1 可用来得到 N_i 的临界值, 即如下定理 2:

定理 2 令 $0 < \beta < 1$,

$$\begin{cases} l_i = \sqrt{\frac{\varphi(1)}{3}} \sigma \sqrt{-2 \log \frac{\beta}{(m-i+1)m}}, \\ R_T(\beta) = \bigcup_{i=1,2,\dots,m} (N_i > l_i), \end{cases} \quad (10)$$

在原假设 H_0 成立下, 若多分辨水平 j 满足式 (7), 则

$$\lim_{T \rightarrow \infty} P\{R_T(\beta)\} \leq \beta. \quad (11)$$

注 1 如果所讨论模型的噪声项具有尺度因子 $\lambda > 0$, 即 $Y_i = f(i/T) + \lambda \epsilon_i, i = 1, \dots, T$, 则式 (10) 中的临界值应为 λl_i . 由定理的证明可得, 临界值 l_i 实际上是由 $w_{j,k}(\epsilon)$ 的极限分布决定的, 也即噪声的不同将导致不同的临界值.

注 2 关于多分辨水平 j 的选定可以比较各个多分辨水平上的 $\max_k |w_{j,k}|$ 与 $C_T = \lambda \sqrt{2\delta} \frac{T^{\delta-1}}{(\log T)^{\delta-1/2}}$, 当函数光滑时 $\max_k |w_{j,k}| \leq C_T$, 当函数含有变点时 $\max_k |w_{j,k}| \geq C_T$ (可参见文献 [6]). 这里 δ 是大于 1 小于 2 的常数, δ 取值的不同会影响到多分辨水平 j 的选取, 在实际应用中可视具体情况而定.

由定理 1、定理 2, 可得如下检验步骤:

步骤 1 选定多分辨水平 j , 对数据做离散小波变换得到经验小波系数, 并进行排序 $|w_{j,(1)}| > |w_{j,(2)}| > \cdots > |w_{j,(2^j)}|$.

步骤 2 令 $N_1 = |w_{j,(1)}| - |w_{j,(m+1)}|, i = 1, 2, \dots, m$. 当检验统计量 $N_1 \leq l_1$ 时, 接受原假设, 即函数 f 为光滑函数; 若 $N_1 > l_1$, 则进行下一步.

步骤 3 比较 N_i 与 $l_i, i = 2, \dots, m$, 若存在常数 p 使得 $N_p > l_p$ 且 $N_{p+1} \leq l_{p+1}, 2 \leq p \leq m$, 则认为函

数 f 含有 p 个变点.

若由以上步骤已知函数含有 p 个变点, 则 $C = l_p + w_{j,(m+1)}$ 为本文选定的小波系数阈值, 即大于此阈值的小波系数对应着函数的跳跃点.

4.2 检验的一致性(The consistency of test)

本节研究在备择假设 H_1 成立时, 变点检验的一致性.

定理 3 令 $R_T(\beta)$ 是定理 2 所定义, 在备择假设 $H_1(m)$ 成立时, 若多分辨水平 j 满足式 (7), 则

$$\lim_{T \rightarrow \infty} P\{R_T(\beta)\} = 1. \quad (12)$$

4.3 变点位置的估计(The estimation of change-points)

令 i_1 为函数 f 的变点个数, 变点位置依次为 $0 < \theta_1 < \theta_2 < \cdots < \theta_{i_1}, 1 \leq i_1 \leq m$, 取 $\hat{i} = \sup(i : 1 \leq i \leq m, T_i > c_i), k_1 < k_2 < \cdots < k_i$ 为 $|w_{j,(1)}|, \dots, |w_{j,(i)}|$ 在未排序时所对应的指标, 不妨设 $|w_{j,k_1}| > |w_{j,k_2}| > \cdots > |w_{j,k_i}|$, 并令 $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_i) = (k_1/2^j, k_2/2^j, \dots, k_i/2^j)$. 由定理 3 可得如下推论:

推论 1 在定理 3 的条件下, 当 $T \rightarrow \infty$ 时, 有

$$P(\hat{i} = i_1) = 1, \quad \sum_{i=1}^{\hat{i}} (\hat{\theta}_i - \theta_i)^2 = O_P(2^{-2j}).$$

由推论 1 及式 (7) 可知, $\hat{\theta}_i - \theta_i = O(2^{-j}) = O(T^{-\delta}(\log T)^\delta)$, 即收敛速度为 $T^{-\delta}(\log T)^\delta$.

5 模拟实验与实例分析(Simulation and real data analysis)

本节用模拟研究与实例分析来说明本文所提方法在变点检测中的有效性.

5.1 模拟实验与分析(Simulation and analysis)

例 1 考虑含有无变点的模型:

$$Y_i = f(i/T) + \lambda \epsilon_i, \quad i = 1, 2, \dots, T,$$

其中 $f(x) = 4 \sin(4\pi x)$, 见图 1(a). 图 1(b) 是噪声过程为 $\epsilon_i = \epsilon_{i-1} + u_i, u_i \sim N(0, 1), i = 1, 2, \dots, T$ 时 Y_i 的显示图, 其中 $\lambda = 0.25$.

取样本容量 $T = 2^{10}$, 小波函数为 Haar 小波, $m = 3$, 检验水平为 $\beta = 0.05$, 取 $\delta = 1.3$, 由注 2 选定多分辨水平 $j = 5$, 在此水平上对图 1(b) 中的 Y_i 进行小波分解, 得到的结果为图 1(c). 由图可知, 所有的小波系数都小于本文给出的阈值 C , 即检验的结果是接受原假设: 函数 f 为光滑函数.

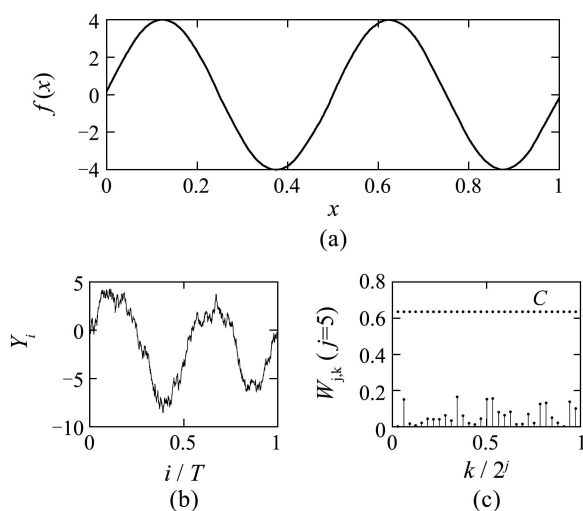


图1 例1的图形

Fig. 1 Figure of example 1

例2 考虑含有两个变点的模型:

$$Y_i = f(i/T) + \lambda \epsilon_i, \quad i = 1, 2, \dots, T,$$

其中 $f(x) = 9 \sin(4\pi x) - 4.5 \operatorname{sgn}(x - 0.3) - 4 \operatorname{sgn}(0.58 - x)$, 见图2(a). 图2(b)是噪声过程为 $\epsilon_i = \epsilon_{i-1} + u_i, u_i \sim N(0, 1), i = 1, 2, \dots, T$ 时 Y_i 的显示图, 其中 $\lambda = 0.25$.

取样本容量 $T = 2^{10}$, 小波函数为Haar小波, $m = 5$, 检验水平为 $\beta = 0.05$. 取 $\delta = 1.3$, 由注2可得只有多分辨水平 $j = 5$ 时才满足要求, 在此水平上对图2(b)中的 Y_i 进行小波分解, 得到的结果为图2(c).

从图中可以看出, 仅有两个小波系数都大于本文给出的阈值 C , 即检验的结果是接受两个变点的备择假设, 与原函数吻合. 同时可得到变点位置的估计值 $\hat{\theta}_1 = 0.3022$ 和 $\hat{\theta}_2 = 0.5906$.

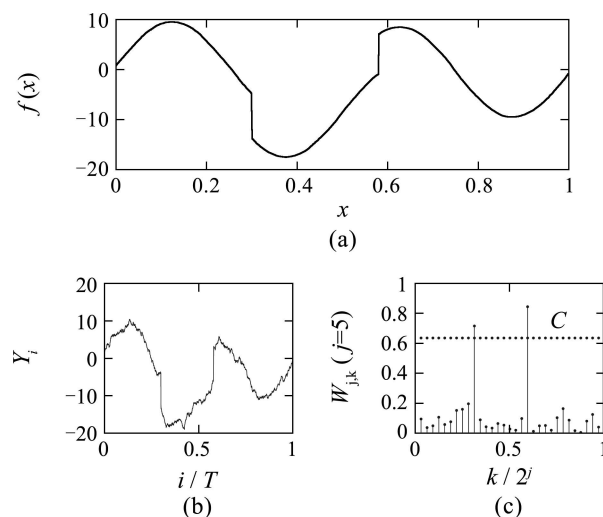


图2 例2的图形

Fig. 2 Figure of example 2

下面将本文的方法与Wang(1995)提出的“UNI”方法(“universal-method”)以及 Odgen, Parrzen(1996)提出的“GOF”方法作比较, 当信噪比取1, 2, 4时分别进行1000次独立的试验, 用Monte-carlo方法近似计算拒绝原假设的概率(详见表1). 这里信噪比指的是 f 的标准差与 ϵ_i 的近似标准差之比.

表1 拒绝 H_0 的Monte-Carlo近似Table 1 Monte-Carlo approximations to the probability of rejecting H_0

信噪比	不含变点			含有变点		
	本文方法	GOF	UNI	本文方法	GOF	UNI
4	0.000	0.000	0.003	0.000	0.931	0.870
2	0.000	0.004	0.031	0.000	0.852	0.792
1	0.000	0.104	0.552	0.992	0.476	0.429

从表1可以看出在原假设下(也即函数不含变点时), 当信噪比较大时3种方法都有较好的效果; 但是当信噪比较低时($\operatorname{snr} = 1$), “UNI”方法和“GOF”方法分别以0.552和0.104的概率拒绝了原假设, 即犯第一类错误的概率较大. 在备择假设下(函数含有变点), 当信噪比较大时本文方法和“GOF”方法检测效果都比较好, 以较大的概率拒绝原假设; 而“UNI”方法即使在信噪比较大的情况下检验效果也不是很好(拒绝原假设的概率分别为0.870和0.792); 当信噪比

降低为1时, “UNI”方法和“GOF”方法分别以0.571和0.524的大概率接受函数不含变点的原假设, 而本文的方法有较好的检测效果.

5.2 实例分析(Real data analysis)

例3 飞机飞行试验是飞机设计、地面试验、设计技术指标、适航性和使用性能的总验证. 试飞测试是飞行试验工作的重要组成部分, 为新机的定型提供了可靠的数据依据. 科研试飞是一项很复杂的系统工程, 在试飞过程中飞机可能会受到系统误差、机器误差、噪声的影响而产生一

些数据跳点, 准确估计数据中跳点发生的位置对新机设计指标和安全性的鉴定有重要的意义^[13]. 本文用一组试飞数据来说明所提方法的实际应用价值.

测试数据的样本容量为2048, 即 $T = 2048$, 对数据作预处理并作图可得图3(a), 从图中可以看出数据是非平稳的, 而且发生了两次跳跃. 对图3(a)的数据在各个多分辨水平上进行小波分解, 取 $\delta = 1.3$, 由注2可得 $j = 7$ 或 $j = 8$ 都满足要求, 这里取 $j = 7$, 再对这个水平上的小波系数进行排序, 求得统计量 N_i 和临界值 l_i , 并依照本文的方法求得小波系数的阈值 C , 结果见图3(b). 从图中不难看出, 仅有两个小波系数超出本文给定的阈值, 这两个小波系数对应着数据的跳跃点, 得到变点位置的估计值 $\hat{\theta}_1 = 0.4423$ 和 $\hat{\theta}_2 = 0.6563$, 也就是说数据在第905个观察值和第1344个观察值处发生跳跃.

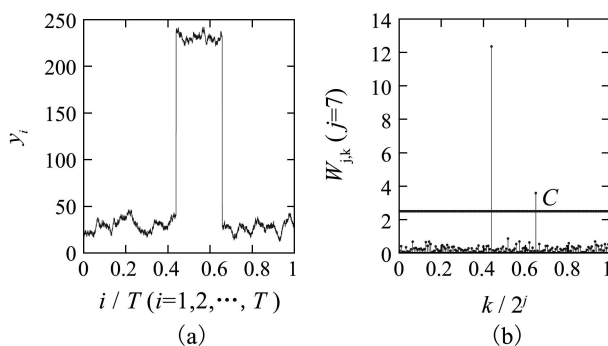


图3 例3的图形

Fig. 3 Figure of example 3

由以上的模拟研究和实例分析可以看出, 利用本文的方法, 可以正确判断出变点的个数并准确的估计变点发生的时刻.

参考文献(References):

- [1] KRISHNAIAH P R, MIAO B Q. Review about estimation of change-point[J]. *Handbook of Statistics*, 1988, 7: 375 – 402.
- [2] 谭常春. 变点问题的统计推断及其在金融中的应用[D]. 合肥: 中国科学技术大学, 2007.
(TAN Changchun. *Statistical inference in change-point and its application in finance*[D]. Hefei: University of Science and Technology of China, 2007.)
- [3] CSÖRGO M, HORVÁTH L. *Limit Theorems in Change-Point Analysis*[M]. West Sussex: Jhon Wiley and Sons Ltd, 1997.
- [4] BAI J. Least squares estimation of a shift in linear processes[J]. *Journal of Time Series Analysis*, 1994, 15(5): 453 – 472.
- [5] KOLOSZKA P S, LEIPUS R. Change point in the mean of dependent observations[J]. *Statistics and Probability Letters*, 1998, 40(4): 385 – 393.
- [6] WANG Y. Jump and sharp cusp detection by wavelet[J]. *Biometrika*, 1995, 82(2): 385 – 397.
- [7] ODGEN T, PARZEN O. Change-point approach to data analytic thresholding[J]. *Statistics and Computing*, 1996, 6(2): 93 – 99.
- [8] RAIHONDO M, TAJVIDI N. A peaks over threshold model for change-point detection by wavelet[J]. *Statistica Sinica*, 2004, 14(2): 395 – 412.
- [9] MIAO B Q, SUBRAMANYAM K. *Some methods to estimate the number and location of slope change points, technical report No.1988, 8*[R]. Pittsburgh: Center for Multivariate Analysis, University of Pittsburgh, 1988.
- [10] 陈懋祖. 高等时间序列计量经济学[M]. 上海: 上海科学技术出版社, 1998.
(CHEN Maozun. *Advanced Time Series Econometrics*[M]. Shanghai: Shanghai Science and Technology Press, 1998.)
- [11] HÄRDLE W, KERKYACHARIAN G, PICARD D, et al. Wavelets, approximation, and statistical application[M] // *Lecture Notes in Statistics*. New York: Springer-Verlag, 1998, Vol.129.
- [12] RAIMONDO M. Minimax estimation of sharp change points[J]. *The Annals of Statistics*, 1998, 26(4): 1379 – 1397.
- [13] 何红丽, 任朴舟, 左益宏. 动态数据异值检测与修复[C] // 中国航空学会测试专业委员会学术年会. [S.l.]: [s.n.], 2004.
(HE Hongli, REN Puzhou, ZUO Yihong. Detection and improvement of outliers to the dynamic data[C] // *Annal Meeting in Testing Speciality Committee of CSAA*. [S.l.]: [s.n.], 2004.)

作者简介:

齐培艳 (1979—), 女, 西北工业大学理学院应用数学系理学硕士, 太原科技大学数学系助教, 主要研究方向为非线性时间序列与信息处理, E-mail: peipei0809@gmail.com; qpy0809@mail.nwpu.edu.cn;

田 铮 (1948—), 女, 教授, 西北工业大学博士生导师, 主要从事非线性时间序列与信息处理、多尺度随机模型与图像处理以及遥感图像统计处理等方面的研究, E-mail: zhtian@nwpu.edu.cn;

段西发 (1979—), 男, 西北工业大学理学院应用数学系理学硕士, 太原科技大学数学系助教, 主要研究方向为信息处理中的快速与并行算法, E-mail: xfaduan@gmail.com;

李晓燕 (1984—), 女, 西北工业大学理学院应用数学系理学硕士, 主要研究方向为非线性时间序列与信息处理, E-mail: xiaoyan18884@126.com;

陈占涛 (1982—), 男, 西北工业大学理学院应用数学系理学博士, 主要研究方向为非线性时间序列与信息处理, E-mail: chenzhanshou@126.com.