第 28 卷第 3 期
2011 年 3 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 28 No. 3
Mar. 2011

文章编号: 1000−8152(2011)03−0351−07

非参数回归模型均值函数结构变点的检测与应用

田　铮[1,2], 赵春辉[1], 陈占寿[1]

(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)

摘要: 本文将一类系统参数变点检测问题转化为非参数回归模型均值函数结构变点的检测问题. 针对当非参数模型均值函数跃度的长期均值为零时, 残量累积和(cumulative sum, CUSUM)统计量无效的问题, 首先利用均值函数的核估计构造新统计量, 给出了原假设和备择假设下统计量的极限分布; 进一步构造Bootstrap检验, 证明了Bootstrap检验的一致性; 最后以模拟结果表明新方法明显优于已有的方法, 并应用于两类实际数据分析, 说明方法的有效性.

关键词: 非参数模型; 均值函数; 结构变点; 核估计; Bootstrap检验.

中图分类号: O23, O29　　**文献标识码:** A

Detection and applications of structural breaks of mean function in nonparametric regression models

TIAN Zheng[1,2], ZHAO Chun-hui[1], CHEN Zhan-shou[1]

(1. Department of Applied Mathematic, Northwestern Polytechnical University, Xi'an Shaanxi 710072, China;
2. State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Applications,
Chinese Academy of Sciences, Beijing 100101, China)

Abstract: The detection of parametric change is transformed into the detection of structural breaks of mean function in the nonparametric models. For the residual cumulative sum(CUSUM) test becomes invalid when the long rang average of jump of the mean function is zero, a new statistic is built based on the kernel estimation of the mean function, and the limiting distributions of null hypothesis and alternative hypothesis are obtained. A Bootstrap procedure is proposed and the consistency of the test is also proved. Finally, simulation and real data analysis are performed to investigate the finite sample properties of our approach. Results show that our method is more powerful than methods proposed in reference.

Key words: nonparametric model; mean function; structural break; kernel estimation; Bootstrap test

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, 质量超过警界线的时刻称为变点. 变点问题在工业自动控制、金融、医学、气候等领域都有明确的实际应用背景[1]. 在工程实际应用中, 系统的异常波动对系统的安全以及产品质量是十分关键的问题, 实质上, 以概率统计观点来看系统的变点检验问题可归结为系统所相应的数学模型的参数变点检测问题.

众所周知, 线性回归模型为刻画变量之间的关系提供了结构简单的数学模型, 与此同时, 线性回归模型回归系数变点的检测问题也得到了大量的研究[2~5]. 但是, 随着科学技术和工程应用的发展, 线性回归模型以及通用的时间序列模型难以精确描述复杂系统中变量的性能, 如金融工程和气象工程中变量的非线性、多尺度等特征, 而一定的统计意义下, 非参数回归模型可逼近具有复杂结构形式系统的观测数据, 在上述工程数据处理中具有明显的优势. 因此, 非参数回归模型变点检测问题受到学者的关注. 如文献[6]研究了固定设计下非参数回归模型均值函数结构变点的检验问题, 文献[7]构造残量累积和(CUSUM)型统计量对随机设计下的非参数回归模型均值函数的结构变点进行检验, 并指出当均值函数跃度的长期均值(long run average) $\lim\limits_{n\to\infty}\dfrac{1}{n}\sum\limits_{i=1}^{n}\mathrm{E}(\Delta(X_i))$为零时残量CUSUM统计量无效, 为此, 文献[7]通过加权来解决这一问题.

但是应注意, 就作者检索国内外最新文献至今尚没有给出最优权函数的选取原则, 当加权后均值函数跃度的长期均值为零时, 对应的加权的CUSUM型统计量依然是无效的. 为此, 本文首先将一类系统的参数变点问题转化为非参数模型均值函数结构变点问题, 基于均值函数的核估计构造Bootstrap方法对非参数回归模型均值函数的结构变点进行检测, 改进了均值函数跃度的长期均值为零时的检验问题, 通过模拟计算和实例分析与已有方法对比, 说明了

收稿日期: 2009−10−23; 收修改稿日期: 2010−04−15.

基金项目: 国家自然科学基金资助项目(10926197, 60972150); 西北工业大学科技创新基金资助项目(2007KJ01033).

本文方法的有效性.

2 问题描述(Problem statement)

考虑如下系统:

$$Y_t = m(X_t, \theta_t) + \sigma_t(X_t)\varepsilon_t,\ t = 1, 2, \cdots, n, \quad (1)$$

其中: $X_t \in \mathbb{R}^d$, $Y_t \in \mathbb{R}^1$为系统可测量, $m(\cdot)$为未知函数, θ为系统参数, ε_t为零均值、不相关随机噪声序列(不可测).

由于θ通常是不可测的、$m(\cdot)$未知和随机扰动, 对θ是否发生变化、何时变化的判断就成为难以处理的问题, 本文对θ的变化进行检测, 即如下假设检验问题:

$$\begin{cases} H_0 : \theta_t = \theta_0,\ t = 1, 2, \cdots, n, \\ H_{\rm A} : \exists \tau_0 \in F,\ k_0 = \lfloor n\tau_0 \rfloor, \\ \theta_t = \theta_0,\ t = 1, 2, \cdots, k_0, \\ \theta_t = \theta_0^* \neq \theta_0,\ t = k_0 + 1, \cdots, n, \end{cases} \quad (2)$$

其中F为$(0, 1)$的闭子区间.

若记$m(x) = m_0(x, \theta_0), \Delta(x) = m_0(x, \theta_0^*) - m_0(x, \theta_0)$, 则上述问题可归结为以下非参数回归模型均值函数结构变点问题:

$$\begin{cases} Y_t = m_t(X_t) + \sigma_t(X_t)\varepsilon_t,\ t = 1, 2, \cdots, n, \\ H_0 \colon m_t(x) = m_0(x),\ t = 1, 2, \cdots, n, \\ H_{\rm A} \colon \exists \tau_0 \in F,\ k_0 = \lfloor n\tau_0 \rfloor, \\ m_t(x) = m_0(x),\ t = 1, 2, \cdots, k_0, \\ m_t(x) = m_0(x) + \Delta(x),\ t = k_0 + 1, \cdots, n, \end{cases} \quad (3)$$

其中F为$(0, 1)$的闭子区间.

注 1 显然假设检验问题(2)包含于问题(3), 问题(3)中, 令$m_t(x) = m_0(x, \theta_0) + \Delta(x)(\theta - \theta_0)$, 则问题(3)包含问题(2), 故问题(2)与问题(3)是等价的.

由于在实际问题中所研究的系统是处于动态稳定中, 对系统的可测量X_t, Y_t做以下假设:

A1) $\{X_t, Y_t\}$ 为强混合序列, 其混合系数为$\alpha(t)$, 且存在$\beta > 2$使$\sum\limits_{t\in N} \alpha(t)^{\beta} < \infty$, $\mathrm{E}Y_t^{\beta} < C < \infty$.

A2) 当$n \to \infty$时, $\dfrac{1}{n}\mathrm{var}(\sum\limits_{t=1}^{n} Y_t) \to \sigma^2 > 0$.

A3) X_t的长期平均分布

$$\bar{F}(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} F_i(x)$$

存在, 其中$F_i(x)$为X_t的分布函数.

A4) $m_0(x)$和$\Delta(x)$在其支撑上满足Lipschitz条件, $\sigma_t(\cdot)$在其支撑上有界.

为研究检验统计量的渐近性质对核函数$K(\cdot)$及窗宽做以下假设:

A5) 当$n \to \infty$时, $h \to 0$, $nh^{2d} \to \infty$.

A6) $K(\cdot)$为$\mathbb{R}^d$上以$\overset{d}{\underset{i=1}{\otimes}}[-1, 1]$为支撑的连续对称核函数.

注 2 A1)∼A3)弱于文献[7]中的假设条件, A4)∼A6)为核估计的一般假设, A1)∼A6)均为非参数函数核估计一致性的充分条件. 由A4)可以看出, 本文只要求条件方差函数有界, 而对其是否随时间变化不做要求.

残量CUSUM统计量利用残量中的均值函数跃度的均值信息对变点进行检验, 故当均值函数跃度的长期均值为零时残量CUSUM统计量是无效的, 但任何一个非零随机变量的一阶绝对矩恒不为零, 本文给出如下基于均值函数核估计的检验统计量:

$$T = \max_{\tau\in F}\{T_\tau\}. \quad (4)$$

其中:

$$T_\tau = \frac{1}{n}\left(\tau\left(1-\tau\right)\right)^{1/2} \sum_{i=1}^{n} \left|m_\tau\left(x_i\right) - m_\tau^*\left(x_i\right)\right|, \quad (5)$$

$$\begin{cases} m_\tau\left(x\right) = \dfrac{\sum\limits_{i=1}^{\lfloor n\tau \rfloor} K_h\left(x - x_i\right) y_i}{\sum\limits_{i=1}^{\lfloor n\tau \rfloor} K_h\left(x - x_i\right)}, \\ m_\tau^*\left(x\right) = \dfrac{\sum\limits_{i=\lfloor n\tau \rfloor+1}^{n} K_h\left(x - x_i\right) y_i}{\sum\limits_{i=\lfloor n\tau \rfloor+1}^{n} K_h\left(x - x_i\right)} \end{cases} \quad (6)$$

均为均值函数$m\left(x\right)$的核估计量.

备择假设下变点分位数的估计量

$$\hat{\tau} = \arg\max_{\tau\in F}\{T_\tau\}. \quad (7)$$

3 非参数回归模型变点的检测(Detection of structural break in nonparametric regression models)

本节研究非参数回归模型均值函数结构变点检验统计量的渐近性质, 构造Bootstrap检验, 证明了检验的一致性.

定理 1 原假设H_0下, 若A1)∼A6)成立, 则

$$T \xrightarrow{d} 0. \quad (8)$$

定理 2 备择假设$H_{\rm A}$下, 若A1)∼A6)成立, 则

$$\hat{\tau} - \tau_0 = \mathrm{O}_p(h^d) + \mathrm{O}_{\rm p}(n^{-1/2}h^{-d}) \to 0, \quad (9)$$

其中$\hat{\tau} = \arg\max\limits_{\tau\in F}\{T_\tau\}$.

推论 1 备择假设$H_{\rm A}$下, 若A1)∼A6)成立, 则

$$T \xrightarrow{d} (\tau_0(1-\tau_0))^{1/2} \int |\Delta(x)|\mathrm{d}\bar{F}(x) > 0. \quad (10)$$

基于极限理论的检验在有限样本下的检验势往往比较差, 非参数检验尤其如此, 由以上定理可以

看出统计量收敛于退化分布, 对应水平下的临界值无法确定, 为了有效地确定在一定检验水平下的拒绝域并提高有限样本下检验的效果, 本文构造如下Bootstrap检验:

步骤 1

步骤 1.1　序列$\{X_i, Y_i\}_{i=1}^n$预处理, 对$\{X_i\}_{i=1}^n$的各维数据乘以相应的系数, 使其各维均为样本方差为1的序列, 得到$\{x_i, y_i = Y_i\}_{i=1}^n$.

步骤 1.2　对核函数$K(\cdot)$, 对序列$\{x_i, y_i = Y_i\}_{i=1}^n$选取带宽

$$h_0 = \arg\min_h \{\sum_{i=1}^n (y_i - m_{-i}(x_i))^2\},$$

其中

$$m_{-i}(x_i) = \frac{\sum\limits_{j\neq i} K((x_i - x_j)/h) y_j}{\sum\limits_{j\neq i} K((x_i - x_j)/h)}.$$

步骤 2

步骤 2.1　对序列$\{x_i, y_i\}_{i=1}^n$计算T和

$$\hat{\tau} = \arg\max_{\tau\in F}\{T_\tau\}.$$

步骤 2.2　计算

$$\{\hat{m}(x_i) = (m_{\hat{\tau}}(x_i) + m^*_{\hat{\tau}}(x_i))/2\}_{i=1}^n.$$

步骤 2.3　计算$\{\hat{U}_i = y_i - \hat{m}(x_i)\}_{i=1}^n$.

步骤 2.4　计算$\bar{\hat{U}} = n^{-1}\sum\limits_{i=1}^n \hat{U}_i$和

$$\{U_i = \hat{U}_i - \bar{\hat{U}}\}_{i=1}^n.$$

步骤 3

步骤 3.1　生成零均值单位方差的独立序列$\{\eta_i\}_{i=1}^n$.

步骤 3.2　计算$\{U_i^* = U_i\eta_i\}_{i=1}^n$和

$$\{y_i^* = \hat{m}(x_i) + U_i^*\}_{i=1}^n.$$

步骤 3.3　类似于$\{x_i, y_i\}_{i=1}^n$计算T, 对于序列$\{x_i, y_i^*\}_{i=1}^n$计算T^*.

步骤 4　重复步骤3 B次, 得到序列 $\{T_i^*\}_{i=1}^B$, 取

$$p^* = \frac{1}{B}\sum_{i=1}^B 1(T \leqslant T_i^*),$$

检验水平为α时, 若$p^* < \alpha$, 则拒绝原假设.

推论 2　在原假设H_0或备择假设$H_{\rm A}$下, 若A1)~A6)成立, 则$T^* \xrightarrow{d} 0$.

注 3　在步骤2和步骤4中, 在有限样本下, 对$\forall\tau\in F$, $\forall t_1 < n\tau (t_2 > n\tau)$, 若

$$\sum_{t=\lfloor n\tau\rfloor}^n 1(x_{t_1} - x_t < h) = 0,$$

$$\sum_{t=1}^{\lfloor n\tau\rfloor} 1(|x_{t_2} - x_t| < h) = 0,$$

则取

$$m_\tau(x_{t_1}) = m^*_\tau(x_{t_1}),\ m^*_\tau(x_{t_2}) = m_\tau(x_{t_2}).$$

注 4　由推论1可以看出这种检验方法只受均值函数变化量的长期一阶绝对矩影响, 而与其长期均值无关, 而随机变量的非零泛函$\Delta(x)$的一阶绝对矩恒不为零, 故由推论1和推论2可知本文给出的Bootstrap检验是一致的, 可以很好的解决基于残量的CUSUM统计量在均值函数的变化量的期望很小时检验效果很差的问题, 避免了权函数的选取问题.

4　模拟结果与实例分析(Simulation and real data analysis)

本节通过模拟计算研究有限样本下跃度的长期均值、长期一阶绝对矩、核函数和窗宽对检测的影响, 对西安气温、日照和上证统计数据进行分析, 说明本文方法的有效性.

非参数模型均值函数核估计中, 窗宽的选择十分关键, 本文采用CV准则[8,9]来选取窗宽, 即

$$h_0 = \arg\min_h \{\mathrm{CV}(h)\} = \arg\min_h \{\sum_{i=1}^n (y_i - m_{-i}(x_i))^2\},$$

其中

$$m_{-i}(x_i) = \frac{\sum\limits_{j\neq i} K((x_i - x_j)/h) y_j}{\sum\limits_{j\neq i} K((x_i - x_j)/h)},$$

$K(\cdot)$为核函数.

文献[8]指出无变点条件下以h_0为窗宽进行的核估计是渐近最优的, 且h_0的选取可以完全由数据驱动, 因而在核估计中得到广泛采用, 模拟结果表明非参数模型均值函数结构变点的检测中利用CV准则选取窗宽是可行的.

4.1　模拟结果(Simulation results)

考虑如下数据生成过程:

$$Y_t = X_t^2 + \alpha\Delta(X_t)1(t > \tau_0 n) + U_t,$$
$$\mathrm{DGP\ 1:}\ \Delta = 1,\ \mathrm{DGP\ 2:}\ \Delta(x) = x,$$

其中X_t和U_t均服从标准正态分布.

例 1　跃度长期均值和长期一阶绝对矩对检测的影响.

对DGP 1和DGP 2分别取$\alpha = 0.5, 1, 2, 0,\ \tau_0 = 0.25, 0.5, 0.75,\ n = 100, 200,\ K(x) = \frac{3}{4}(1 - x^2)1(|x| < 1),\ B = 500$. 通过Bootstrap方法对变点进行检验, 重复200次实验, 记录拒绝原假设的次数, 5%名义水平下, 拒绝原假设(不存在均值函数结构变点)的频率见表1.

表 1 本文方法的检验功效(检验水平0.05)

Table 1 Power of our test (nominal level: 0.05)

n	DGP	τ_0	α			
			0.5	1	2	0
100	1	0.25	0.245	0.705	1	0.04
		0.5	0.31	0.875	1	
		0.75	0.195	0.705	1	
	2	0.25	0.145	0.45	0.955	
		0.5	0.16	0.555	0.995	
		0.75	0.145	0.44	0.985	
200	1	0.25	0.56	0.985	1	0.04
		0.5	0.55	0.985	1	
		0.75	0.54	0.99	1	
	2	0.25	0.22	0.75	1	
		0.5	0.32	0.95	1	
		0.75	0.20	0.775	1	

为研究样本量、变点分位数、均值函数跃度的长期一阶绝对矩和长期均值对变点分位数估计的影响, 对DGP 1和DGP 2分别取$\tau_0 = 0.25, 0.5, 0.75$, 均值函数跃度的一阶绝对矩

$$d = \frac{1}{n}\mathrm{E}(\sum_{t=1}^{n} \alpha * |\Delta(X_t)|) = 1, 2,\ n = 100, 200,$$

对变点分位数进行估计, 重复200次实验, 变点分位数估计值的样本均值和标准差见表2(括号内的为标准差).

表 2 有限样本下变点分位数的估计

Table 2 The estimation of change quartile

n	DGP	τ_0	d	
			1	2
100	1	0.25	0.2637(0.1304)	0.2403(0.0216)
		0.5	0.4744(0.1081)	0.4885(0.0206)
		0.75	0.7045(0.1368)	0.7398(0.0175)
	2	0.25	0.2783(0.1230)	0.2540(0.0378)
		0.5	0.4915(0.1026)	0.4905(0.0240)
		0.75	0.7022(0.1075)	0.7340(0.0284)
200	1	0.25	0.2427(0.0372)	0.2437(0.0096)
		0.5	0.4906(0.0544)	0.4945(0.0084)
		0.75	0.7455(0.0515)	0.7448(0.0096)
	2	0.25	0.2539(0.0444)	0.2486(0.0150)
		0.5	0.4935(0.0423)	0.4965(0.0081)
		0.75	0.7336(0.0622)	0.7418(0.0122)

表1中结果表明: 原假设下本文方法能够正确地判断模拟数据不存在均值函数结构变点; 在备择假设下, 本文的方法具有良好的功效, 随着样本量的增加或跃度的长期一阶绝对矩增大, 检验的势随之增加; 变点越靠近中间位置, 经验势越高. 本文方法受均值函数跃度的长期均值的影响较小, 无论均值函数跃度的长期均值是否为零, 本文的方法均明显的优于已有方法, 有效地改进了非参数回归模型均值函数结构变点的检验.

表2中的结果表明: 变点分位数的估计值偏左; 随着样本量增加或跃度一阶绝对矩的增大, 变点分位数估计随之更准确; 跃度的长期均值对变点分位数的估计影响不大.

例 2 核函数和窗宽对检测的影响.

对数据生成过程 DGP 2 分别取核函数为Epanechnikov核和截断高斯核:

$$K(x) = C\exp(-x^2/2)1(|x| < 1),$$
$$C = \int_{-1}^{1} \exp(-x^2/2)\mathrm{d}x.$$

取窗宽h_n, 为$h_n = h_0 n^{0.2-1/r}(r = 4, 5, 7, 9, 10)$, $B = 500$进行模拟试验, 重复1000次独立实验得到检验和估计结果分别见表3、表4.

注 5 表4中每个表格中第1个数为估计值的均值, 第2个为估计值的标准差.

表3, 4中结果表明: 通过CV准则选取窗宽, 本文方法可以有效的检测非参数模型均值函数结构变点; 随着样本量增加或跃度长期一阶绝对矩的增大, 检验的势随之增大, 估计也更准确; 变点检测受核函数和窗宽选择的影响不大.

4.2 实例分析(Real data analysis)

本节在0.05的检验水平下对两组实际数据进行变点检测.

例 3 1980∼1989年气温数据均值函数结构变点检测.

气候与人类的生活息息相关, 气候的变化对农业[10,11]、生态[12]、人类健康[13]等有重要的影响, 准确检测气候诸因素的变化对农业结构的调整、病虫害的防治、生态环境的保护、传染病的防控有着重要的意义.

气温是气候的重要因素之一, 主要受日照、时间、人类活动、大气环流、洋流和下垫层的影响, 在一个地区大气环流、洋流和下垫层相对稳定, 人类活动是否对气温构成影响难以观察, 气温受日照和时间的影响较大, 因此可以通过研究温度与日照、时间的关系是否发生变化来判断人类活动是否导致了气温的变化. 本文对西安市1980∼1990年温度、时间和日照的旬值(360个数据)进行研究, 所选数据来自中国气象科学数据共享网(http://cdc.cma.gov.cn/shuju/preview.jsp), 可确保其真实性和准确性.

表 3 采用不同的窗宽和核函数得到的检验结果

Table 3 Test results with series of smooth parameters and kernels

n	τ_0	α	r									
			Epanechnikov核					截断高斯核				
			4	5	7	9	10	4	5	7	9	10
100	0.25	0.5	0.116	0.097	0.087	0.11	0.091	0.112	0.112	0.097	0.069	0.091
		1	0.4	0.402	0.403	0.417	0.396	0.406	0.383	0.382	0.406	0.389
		2	0.955	0.961	0.964	0.972	0.977	0.961	0.963	0.979	0.96	0.96
	0.5	0.5	0.168	0.145	0.156	0.148	0.149	0.171	0.151	0.126	0.13	0.122
		1	0.604	0.623	0.624	0.641	0.602	0.575	0.614	0.616	0.594	0.573
		2	0.994	0.994	0.998	0.99	0.993	0.994	0.995	0.996	0.996	0.998
	0.75	0.5	0.103	0.086	0.081	0.077	0.083	0.103	0.101	0.087	0.08	0.086
		1	0.357	0.371	0.366	0.349	0.375	0.343	0.359	0.356	0.338	0.348
		2	0.934	0.955	0.947	0.957	0.961	0.938	0.957	0.959	0.951	0.959
		0	0.068	0.058	0.03	0.035	0.043	0.052	0.043	0.05	0.05	0.047
200	0.25	0.5	0.219	0.215	0.251	0.242	0.201	0.212	0.231	0.218	0.219	0.216
		1	0.826	0.817	0.836	0.865	0.873	0.824	0.833	0.858	0.827	0.864
		2	1	1	1	1	1	1	1	1	1	1
	0.5	0.5	0.321	0.353	0.331	0.335	0.375	0.333	0.337	0.346	0.317	0.338
		1	0.947	0.96	0.966	0.976	0.967	0.961	0.957	0.978	0.964	0.963
		2	1	1	1	1	1	1	1	1	1	1
	0.75	0.5	0.221	0.201	0.199	0.221	0.198	0.197	0.2	0.197	0.169	0.199
		1	0.81	0.827	0.834	0.851	0.843	0.804	0.823	0.848	0.844	0.838
		2	1	1	1	1	1	1	1	1	1	1
		0	0.043	0.051	0.046	0.052	0.048	0.041	0.038	0.053	0.04	0.036

表 4 采用不同的窗宽和核函数得到的估计结果

Table 4 Estimation results with series of smooth parameters and kernels

n	τ_0	d	r									
			Epanechnikov核					截断高斯核				
			4	5	7	9	10	4	5	7	9	10
100	0.25	1	0.3025	0.2893	0.2892	0.2857	0.2899	0.3069	0.3002	0.2892	0.2951	0.2968
			0.1472	0.1399	0.1489	0.1407	0.1478	0.1607	0.1572	0.1444	0.1502	0.1465
		2	0.253	0.2502	0.2484	0.2483	0.2489	0.2537	0.2507	0.2495	0.2523	0.2502
			0.0345	0.0311	0.0302	0.033	0.0297	0.0376	0.033	0.0293	0.0366	0.0417
	0.5	1	0.4865	0.4916	0.4821	0.4936	0.4919	0.4916	0.4845	0.492	0.4969	0.4957
			0.1068	0.101	0.103	0.0941	0.0987	0.1049	0.1032	0.1013	0.0969	0.1001
		2	0.4916	0.4911	0.4923	0.4929	0.494	0.4923	0.4918	0.4936	0.4944	0.4933
			0.0254	0.0235	0.0204	0.0255	0.0221	0.0224	0.0204	0.0202	0.0236	0.0196
	0.75	1	0.6948	0.6966	0.6985	0.7165	0.6999	0.6917	0.7011	0.7071	0.7132	0.7079
			0.1334	0.1397	0.133	0.1132	0.1391	0.1423	0.1347	0.1247	0.1261	0.1296
		2	0.7314	0.7328	0.7354	0.7352	0.7371	0.7314	0.7325	0.7363	0.7387	0.7388
			0.0269	0.0276	0.0242	0.0257	0.0242	0.0298	0.0334	0.0262	0.0218	0.0216
200	0.25	1	0.2571	0.2528	0.253	0.2534	0.2514	0.2858	0.2689	0.2687	0.2732	0.2694
			0.0805	0.0554	0.057	0.0532	0.0562	0.1022	0.0771	0.0876	0.092	0.0897
		2	0.2485	0.249	0.25	0.2487	0.2486	0.2553	0.2547	0.2562	0.2547	0.2548
			0.0144	0.0131	0.0139	0.0131	0.0131	0.0203	0.0191	0.0185	0.0156	0.0158
	0.5	1	0.493	0.495	0.4952	0.4957	0.496	0.4966	0.4977	0.4999	0.5013	0.4996
			0.049	0.0481	0.0438	0.0432	0.0375	0.0675	0.0581	0.0548	0.0557	0.0638
		2	0.4948	0.4957	0.496	0.4967	0.4963	0.5003	0.4998	0.5009	0.5014	0.50123
			0.0104	0.0107	0.0103	0.0089	0.0093	0.0155	0.0135	0.0127	0.013	0.0132
	0.75	1	0.7351	0.7394	0.7401	0.742	0.7406	0.719	0.7224	0.726	0.7286	0.7249
			0.0617	0.052	0.0454	0.0379	0.047	0.0895	0.0908	0.0906	0.0866	0.0917
		2	0.7419	0.743	0.7422	0.7434	0.7438	0.7446	0.7454	0.7469	0.7467	0.7474
			0.0126	0.0108	0.012	0.0097	0.0106	0.0186	0.0172	0.0147	0.0153	0.0156

将1∼12月的上中下旬按时间先后对应为1, $\cdots$, 36, 与对应的日照时间(小时)构成二维向量$\{X_t\}$, 温度(0.1°)按时间先后顺序构成序列$\{Y_t\}$.

按照步骤1.1对$\{X_t\}$进行预处理;

依据步骤1.2求得窗宽$h_0 = 0.0406$;

按照注4和步骤2.1计算统计量, 在$\hat{\tau} = 181/360$(1985年1月上旬)处, 统计量达到最大7.0025.

按照步骤2生成序列$\{\hat{m}(x_t)\}_{t=1}^n$和$\{\bar{\hat{U}}\}_{t=1}^n$.

步骤3.1生成独立标准正态分布序列$\{\eta_i\}_{i=1}^n$.

按照步骤3.2, 3.3计算统计量T^*.

步骤4重复步骤3, $B = 500$次, 得到统计量序列 $\{T_b^*\}_{b=1}^B$如图 1 所示. $p^* = \dfrac{1}{B}\sum\limits_{b=1}^{B} 1(T \leqslant T_b^*) = 0/500 < 0.05$, 拒绝原假设, 认为在$\hat{\tau} = 181/360$(1985年1月上旬)处存在变点.

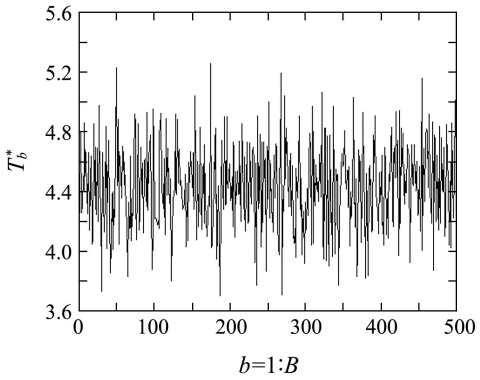

图 1 Bootstrap统计量序列T_b^*

Fig. 1 Bootstrap statistics T_b^*

可以看出1985年1月上旬检测到变点, 说明在这一时间点前后人类的活动明显的导致了温度的变化, 这与1983∼1985年严重的经济过热相对应, 这说明当时我国大规模的建设和短时间内能源消耗量的急剧上升导致了气温的显著变化, 符合温室效应和城市化会导致温度的上升[14]的结论.

例 4 股市2008年7月至2009年9月异常波动的检测.

变点检测可以用来验证某些未被研究的量是否能够反映系统中的某些变化, 从而为更好的对系统监控提供更多的手段.

本文研究 2008 年7月1日到 2009 年9月24日共307个交易日上证交易所的交易数据(http://biz.finance.sina.com.cn/company/history.php). 对证券交易中收盘价(CP_t), 当日交易量(VT_t)和交易额(FT_t)之间的关系知之甚少, 考虑到单笔交易增长率会受到当日股价的增长率的影响, 取

$$X_t = FT_t * VT_{t-1}/VT_t/FT_{t-1},\ Y_t = CP_t/CP_{t-1},$$

对$Y_t = f_t(X_t) + \sigma_t(X_t)\varepsilon_t$均值函数的结构变点进行研究, 没有检测到变点. 按照例1的方法对$\{X_t, Y_t\}$进行变点检测, $h_0 = 0.3184$ 在137处统计量达到最大0.0045, $B = 500$得到的统计量序列

$$p^* = \frac{1}{B}\sum_{b=1}^{B} 1(T \leqslant T_b^*) = 31/500 > 0.05,$$

接受原假设, 认为不存在变点.

研究

$$X_t^* = VT_tCP_{t-1}/VT_{t-1}/CP_t \text{和} Y_t^* = FT_t/FT_{t-1}$$

的关系的文章尚未出现, 但从散点图(图2)可以看出其满足一定的关系. 按照例1的方法对$\{X_t^*, Y_t^*\}$进行变点检测, 237(2009年6月19日)处统计量达到最大0.0194, $B = 500$得到的统计量序列, $p^* = \dfrac{1}{B}\sum\limits_{b=1}^{B} 1(T \leqslant T_b^*) = 1/500 < 0.05$, 拒绝原假设, 认为237(2009年6月19日)处存在变点.

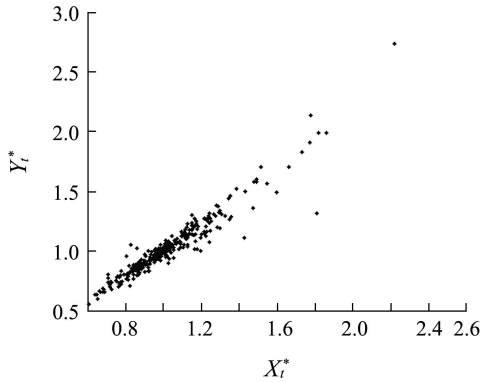

图 2 X_t^*, Y_t^*的散点图

Fig. 2 Scatter of X_t^*, Y_t^*

利用本文的方法对$Y_t^* = f_t(X_t^*) + \sigma_t(X_t^*)\varepsilon_t$均值函数的结构变点进行研究, 发现2009年6月19日为一个变点, 变点前后的散点图分别见图3、图4, 这与财政部、国资委、证监会和全国社保基金理事会2009年6月19日联合发布公告, 在境内证券市场实施国有股转持政策相对应, 说明二者存在某种关系.

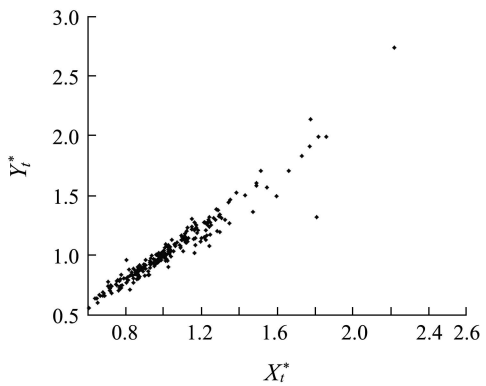

图 3 变点前X_t^*, Y_t^* 的散点图

Fig. 3 Scatter of X_t^*, Y_t^* before the change

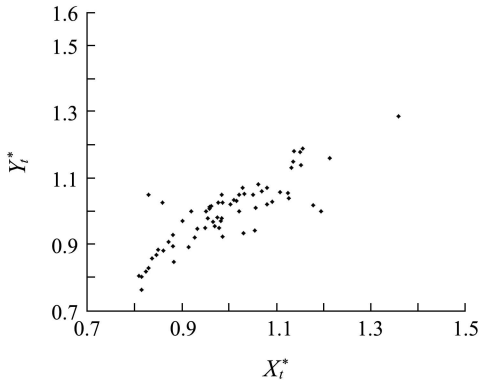

图 4　变点后X_t^*, Y_t^*的散点图

Fig. 4　Scatter of X_t^*, Y_t^* after the change

均值函数跃度的长期均值较小, 文献中的方法不能检测到这个变点, 说明了本文方法比已有方法更有效.

5　结束语(Conclusion)

本文讨论了一类非参数回归模型均值函数结构变点相等价的系统参数变点的检测问题, 实例分析和模拟计算结果均说明本文的方法可以有效地检测到非参数回归模型均值函数的结构变点, 有效地改进了跃度长期平均期望为零时非参数模型均值函数的结构变点的检测.

文中为保证理论结果成立而设置的假设条件比较严格, 且没有关于核估计中窗宽选择的相关理论结果, 更宽松条件下非参数模型均值函数结构变点的检测及窗宽选择对其的影响的相关问题, 我们正在研究.

参考文献(References):

[1] 齐培艳, 田铮, 段西发, 等. 噪声为单位根过程的非参数函数变点的小波检测[J]. 控制理论与应用, 2009, 26(1): 57 – 61.
(QI Peiyan, TIAN Zheng, DUAN Xifa, et al. Wavelet detection of jumping points in a nonparametric function with the unit-root noise[J]. *Control Theory & Applications*, 2009, 26(1): 57 – 61.)

[2] BROWN R L, DURBIN J, EVANS J M. Techniques for testing the constancy of regression relationships over time[J]. *Journal of the Royal Statistical Society, Series B*, 1975, 37(2): 149 – 192.

[3] PLOBERGER W, KRAMER W. The local power of the CUSUM and CUSUM of squares tests[J]. *Econometric Theory*, 1990, 6(3): 335 – 347.

[4] PLOBERGER W, KRAMER W. The CUSUM test with OLS residuals[J]. *Econometrica*, 1992, 60(2): 271 – 285.

[5] RICHARD LUGER. A modifed CUSUM test for orthogonal structural changes[J]. *Economics Letters*, 2001, 73(3): 301 – 306.

[6] LOADER C R. Change point estimation using nonparametric regression[J]. *Annals of Statistics*, 1996, 24(4): 1667 – 1678.

[7] SU L J, XIAO Z J. Testing structural change in time-series nonparametric regression models[J]. *Statistics and Its Interface*, 2008, 1(1): 347 – 366.

[8] WU J S, CHU C K. Nonparametric function estimation and bandwidth selection for discontinuous functions[J]. *Statistica Anica*, 1993, 3: 557 – 576.

[9] IRENE GIJBELS. Anne-cecile goderniaux bandwidth selection for change point estimation in nonparameteric regression[J]. *American Statistical Association and the American Sosiety for Quality Tecnometrics*, 2004, 46(1): 76 – 96.

[10] 王亚平. 1960∼2005年气候变化对东北三省地表干湿状况和作物生长期的影响[D]. 南京: 南京农业大学, 2008.
(WANG Yaping. *The impacts of climate change on surface dry-wet status and growing season from 1960 to 2005 in three provinces of northwest China*[D]. Nanjing: Nanjing Agricultural University, 2008.)

[11] 姚树然, 霍治国, 关福来, 等. 气候及其变化对飞蝗发生期的影响[J]. 生态学杂志, 2009, 28(7) : 1356 – 1360.
(YAO Shuran, HUO Zhiguo, GUAN Fulai, et al. Effects of climate and its change on the occurrence of oriental migratory locust around bohaibay[J]. *Chinese Journal of Ecology*, 2009, 28(7): 1356 – 1360.)

[12] 李兴华, 韩芳, 张存厚, 等. 气候变化对内蒙古中东部沙地、湿地镶嵌景观的影响[J]. 应用生态学报, 2009, 20(1): 105 – 112.
(LI Xinghua, HAN Fang, ZHANG Cunhou, et al. Influence of climate change on mosaic landscape of sand land-wetland in middle-east Inner Mongolia[J]. *Chinese Journal of Applied Ecology*, 2009, 20(1): 105 – 112.)

[13] 褚秀娟, 郭家钢. 气候变暖对血吸虫病传播的影响及相关研究技术的应用[J]. 中国寄生虫学与寄生虫病杂志, 2009, 27(3): 265 – 269.
(CHU Xiujuan, GUO Jiagang. Impact of climate warming on schistosomiasis transmission and application of relative research techniques[J]. *Chinese Journal of Parasitology and Parasitic Diseases*, 2009, 27(3): 265 – 269.)

[14] 王绍武, 叶瑾琳. 近百年全球气候变暖的分析[J]. 大气科学, 1995, 19(5): 545 – 553.
(WANG Shaowu, YE Jinlin. An analysis of global warming during the last one handred years[J]. *Scientia Atmospherica Sinica*, 1995, 19(5): 545 – 553.)

作者简介:

田　铮　(1948—), 女, 教授, 博士生导师, 主要从事非线性时间序列分析与遥感图像信息处理、图理论与SAR图像理解、模式识别与遥感图像统计处理等领域的研究, E-mail: zhtian@nwpu.edu.cn;

赵春辉　(1985—), 男, 硕士研究生, 主要研究金融非线性时间序列分析理论、方法与应用, E-mail: zc3h@sina.com;

陈占寿　(1982—), 男, 博士研究生, 主要研究金融非线性时间序列分析理论、方法与应用, E-mail: chenzhanshou@126.com.