doi:10. 16055/j. issn. 1672-058X. 2020. 0006. 006

基于二元分割检测分段线性趋势中的多变点*

刘 伟[1], 胡 尧[1,2**], 胡 倩[1]

(1. 贵州大学 数学与统计学院, 贵阳 550025; 2. 贵州省公共大数据重点实验室, 贵阳 550025)

摘 要:变点检测问题一直是统计学中的热点研究之一,在实际的数据中, 通常会在某一段具有线性增长或减少的趋势, 这种趋势的起始点位置是未知的,因此针对此种具有分段线性趋势的一维数据, 提出了一种多变点检测方法。该方法根据广义对数似然比所构造出的统计量, 将二元分割方法、阈值准则和 sSIC 三者相结合, 能快速有效地检测出数据中的多变点。数值模拟结果表明, 对具有分段线性趋势的数据, 检测变点的位置及数量很准确,检测结果令人满意。最后以深圳市北环大道新洲立交的车流量数据为例, 分析出该区域在工作日和非工作日的变点分布特征, 分析结果符合实际情况, 可为交管部门的相关工作提供参考意见。

关键词:多变点检测; sSIC; 二元分割; 线性趋势分段

中图分类号:O212 **文献标志码**:A **文章编号**:1672-058X(2020)06-00032-07

0 引 言

变点检测问题一直是统计学中一个经典的分支, 最初起源于 Page 在 Biometrika 上发表的一篇关于连续抽样检验的文章[1], 这是一篇关于质量检测问题的理论研究, 从此开启了变点研究的篇章。在如今的大数据发展趋势中, 掌握数据中的突变对于分析数据, 挖掘其中的隐藏信息有着至关重要的作用, 所以越来越多的统计学者投入到这一研究领域[2-3]。近年来, 变点检测问题已被广泛应用于各个领域, 在许多行业中, 都能看到变点的身影,例如自动检测云数据中的变点[4], 以保持应用程序或网站的性能和可用性; 热带气旋数据中的气候变化检测[5], 能及时预防重大自然灾害;根据光曲线数据的变化能检测系外行星; DNA 拷贝数的突变跟某些疾病的起因密切相关[6]; 对潜在协整股票价格的平稳区间的估计可以降低损失的风险[7]等。

所谓的变点, 就是在一个时间序列或过程中, 当某个统计特性在某一时刻 τ 突然发生了变化, 就称该时刻 τ 为变点, 统计特性包括样本的分布类型、分布参数、数字特征等, 变点检测就是利用统计量或统计方法将该时刻 τ 估计出来。在数据被假定为分段恒定的情况下, 一类常用的方法是基于最小化成本函数的思想, 如 Jackson 等[8]提出的 OP(Optimal Partitioning)算法, 该算法是在成本函数中引入惩罚项, 将变点检测转化为成本函数惩罚最小化问题, 但是在数据量比较大的情况下, 计算比较复杂; 所以 Killick 等[9]提出基于不等式修剪的 PELT(Pruned Exact Linear Time)算法, 它比 OP 更

收稿日期:2020-02-14;**修回日期**:2020-03-15.

* **基金项目**:国家自然科学基金资助项目(11661018); 贵州省科技计划项目(黔科合平台人才[2017]5788 号).

作者简介:刘伟(1994—),男,四川宜宾人, 硕士研究生,从事概率论与数理统计研究.

** **通讯作者**:胡尧(1971—),男, 贵州遵义人, 教授, 硕士生导师, 从事应用统计研究. Email: yhu1@gzu. edu. cn.

有效且计算简单;而 Maidstone 等[10]将 PELT 与 pDPA(pruned Dynamic Programming Algorithm)相结合,提出一种更稳健高效的 FPOP(Functional Pruning Optimal Partitioning)算法等。而在数据具有线性趋势变化的相关研究中,Bai 和 Perron[11]考虑通过最小二乘法估计具有多个结构变化的线性模型,并针对无变化的原假设提出 Wald 型检验;Kim 等[12]和 Tibshirani 等[13]考虑了具有 L_1 惩罚的“趋势过滤”;Fearnhead 和 Maidstone 等[14]通过动态规划算法用 L_0 正则化来检测斜率的变化;Spiriti 等[15]研究了两种优化最小二乘和惩罚样条中节点位置的算法;Anastasiou 和 Fryzlewicz 提出了 ID(Isolation-Detection)方法,该方法不断地搜索扩展的数据段以检测其中变化,但正因如此,会使得某些数据被多次重复计算,而且每次扩展的数据量只给出一个固定值 $\lambda=3$,并没有说明给出的原因,并且在数据为长时间的小跳跃情况下,该方法比较乏力。

二元分割方法[16](Binary Segmentation,BS)是多变点检测的经典方法之一,与其他变点检测方法相比,该方法检测效果很好,特别是对大量数据,长期性数据的多变点检测,很多单变点检测方法都能跟二元分割相结合而转化为多变点检测,如 Olshen A B 等[17]的 CBS(Circular Binary Segmentation)方法,Fryzlewicz 的 WBS(Wild Binary Segmentation)方法[18]和 WBS2(Wild Binary Segmentation 2)方法等,但这两种方法都是用于检测均值变点,所以在数据存在异常值时,检测结果会存在很大偏差。在二元分割方法中,检验统计量非常重要,所以在本文中,根据 Baranowski 等[19]提出的统计量作为的检验统计量,同时受到 WBS 理论对整个数据区域随机“产生”区间以检测变点的启发,也对整个数据序列随机抽取检测区间进行变点检测。

1　模型及检测方法

1.1　模型介绍

在具体应用过程中,由于数据类型的多样性,不同的数据,其分布类型不能确定,参数方法已经无法满足实际应用的需求,然而非参数方法对总体分布的假定要求低,不会因为对总体分布的假定不当而导致重大问题,更能体现让数据说话的特点,具有很好的稳健性,所以基于非参数模型对变点进行研究更具有通用性。

对于观测的数据序列 $Y=(Y_1,Y_2,\cdots,Y_T)$,运用如下经典的单变量统计模型:

$$Y_t=f_t+\sigma_t\varepsilon_t,\ t=1,2,\cdots,T \tag{1}$$

其中,Y_t 为单次观测数据,f_t 是确定的数据信号,ε_t 为独立的随机噪声,且 $\varepsilon_t\sim N(0,\sigma_t^2)$,在第二节数值研究设 $\sigma_t=1$。假设时间序列数据 Y 有 q 个变点,则 Y 被分割为 $q+1$ 个不同的区间段,记变点的位置分别为 $0=\tau_0<\tau_1<,\cdots,<\tau_q<\tau_{q+1}=T$,其中 q 的值和变点位置 τ 未知,需要估计,第 i 个区间段的数据为 $Y_{\tau_{i-1}+1:\tau_i}=(Y_{\tau_{i-1}+1},Y_{\tau_{i-1}+2},\cdots,Y_{\tau_i})$,本文研究的是分段线性趋势的数据,所以此处每个区间段的数据呈线性相关,研究的目的是估计出每个区间段的首尾端点,也就是变点。

在该模型中,设 $\sigma_t=\sigma$,则对 σ 的估计,可以运用中位数绝对偏差(Median Absolute Deviation,MAD)方法[20],在 ε_t 为独立同分布的高斯情况下,MAD 定义为

$$\hat{\sigma}=\frac{median(|Y_1-2Y_2+Y_3|,|Y_2-2Y_3+Y_4|,\cdots,|Y_{T-2}-2Y_{T-1}+Y_T|)}{\Phi^{-1}(3/4)\sqrt{6}}$$

其中,$\Phi^{-1}(\cdot)$ 表示标准正态分布的分位数函数。注意,MAD 的估计值对 f_t 中的任何变点都是稳健的,因为它结合了对差异数据的处理和对中位数的使用。

1.2　检验统计量

检验统计量作为变点识别的主要部分,其检测能力直接影响变点检测的最终结果,选择一个好的检验统计量至关重要,所以选择 Baranowski 在 2019 年所提出的统计量[19]为本文的检验统计量,具体构造如下。

在随机噪声 ε_t 服从高斯分布的条件下,设 $R^b_{(s,e]}(\boldsymbol{v})$ 是区间 $(s,e]$ 中潜在的单个变点的广义对数似然比,$s<b\leqslant e$,向量 $\boldsymbol{v}$ 是 T 维的,$\boldsymbol{v}=(0,\cdots,0,$

$Y_{s+1},Y_{s+2},\cdots,Y_e,\cdots,0)^{\mathrm{T}}$, 构造检验统计量的思想是找到一个对比函数 $C^b_{(s,e]}(\boldsymbol{v})$, 使得

$$\arg\max_{s<b\leqslant e}C^b_{(s,e]}(\boldsymbol{v})=\arg\max_{s<b\leqslant e}R^b_{(s,e]}(\boldsymbol{v})$$

而对比函数是由数据与对比向量的内积所构成, 定义对比函数为

$$C^b_{(s,e]}(\boldsymbol{v})=|\langle\boldsymbol{v},\boldsymbol{\varphi}^b_{(s,e]}\rangle| \tag{2}$$

其中,

$$\alpha^b_{(s,e]}=\left\{\frac{6}{l(l^2-1)[1+(e-b+1)(b-s)+(e-b)(b-s-1)]}\right\}^{\frac{1}{2}}$$

$$\beta^b_{(s,e]}=\left\{\frac{(e-b+1)(e-b)}{(b-s-1)(b-s)}\right\}^{\frac{1}{2}},l=e-s$$

为了解释选择三角函数 $\varphi^b_{s,e}(\cdot)$的原因, 为区间$(s,e]$定义了线性向量:

$$\boldsymbol{\gamma}_{(s,e]}=(\boldsymbol{\gamma}_{(s,e]}(1),\boldsymbol{\gamma}_{(s,e]}(2),\cdots,\boldsymbol{\gamma}_{(s,e]}(T))^{\mathrm{T}},$$

$$\boldsymbol{\gamma}_{(s,e]}(t)=\begin{cases}\left\{\frac{1}{12}(e-s-1)(e-s)(e-s+1)\right\}^{-1/2}\times\\ \qquad\left(t-\frac{e+s+1}{2}\right),t=s+1,s+2,\cdots,e\\ 0,\text{其他}\end{cases}$$

以及常数向量

$$\mathbf{1}_{(s,e]}=\mathbf{1}_{(s,e]}(1),\mathbf{1}_{(s,e]}(2),\cdots,\mathbf{1}_{(s,e]}(\mathrm{T}))^{\mathrm{T}},$$

$$\mathbf{1}_{(s,e]}(t)=\begin{cases}(e-s)^{-1/2},t=s+1,s+2,\cdots,e\\ 0,\text{其他}\end{cases}$$

在向量

$$\tilde{\boldsymbol{\varphi}}^b_{(s,e]}(t)=\begin{cases}t-b,t=b+1,b+2,\cdots,e\\ 0,\text{其他}\end{cases}$$

上($b+1$ 处是一个转折点),对 $\boldsymbol{\gamma}_{(s,e]}(t)$ 和 $\mathbf{1}_{(s,e]}(t)$ 应用 Gram-Schmidt 正交化, 并对得出的向量进行标准化, 使得 $\|\cdot\|_2=1$, 最终得出对比向量 $\boldsymbol{\varphi}^b_{(s,e]}(t)$。对于欧几里德距离而言, $(s,e]$中 Y_t 的最佳逼近是 $\boldsymbol{\gamma}_{(s,e]}(t)$, $\mathbf{1}_{(s,e]}(t)$ 和 $\boldsymbol{\varphi}^b_{(s,e]}(t)$ 的线性组合, 此三者是相互正交的, 正是这种标准正交性使得

对于区间$(s,e]$, 在给定$(Y_{s+1},Y_{s+2},\cdots,Y_e)$的情况下, 设 $l(Y_{s+1},Y_{s+2},\cdots,Y_e;\boldsymbol{\Theta})$为 $\boldsymbol{\Theta}$ 的似然,$\boldsymbol{\Theta}$ 为参数空间, 则广义对数似然比定义如下:

$$R^b_{(s,e]}(\boldsymbol{v})=2\log\frac{\sup_{\Theta^1,\Theta^2}l(Y_{s+1},Y_{s+2},\cdots,Y_b;\boldsymbol{\Theta}^1)l(Y_{b+1},Y_{b+2},\cdots,Y_e;\boldsymbol{\Theta}^2)}{\sup_{\Theta}l(Y_{s+1},Y_{s+2},\cdots,Y_e;\boldsymbol{\Theta})}$$

对于连续的具有分段线性趋势的数据,适当的对比向量是 $\boldsymbol{\varphi}^b_{(s,e]}=(\boldsymbol{\varphi}^b_{(s,e]}(1),\boldsymbol{\varphi}^b_{(s,e]}(2),\cdots,\boldsymbol{\varphi}^b_{(s,e]}(T))^{\mathrm{T}}$, 其中

$$\boldsymbol{\varphi}^b_{(s,e]}(t)=\begin{cases}\alpha^b_{(s,e]}\beta^b_{(s,e]}[(e+2b-3s-1)t-(be+bs-b-2s^2-2s)],t=s+1,s+2,\cdots,b\\ -\frac{\alpha^b_{(s,e]}}{\beta^b_{(s,e]}}[(3e-2b-s+1)t-(2e^2+2e-be-bs-b)],t=b+1,b+2,\cdots,e\\ 0,\text{其他}\end{cases} \tag{3}$$

$$R^b_{(s,e]}(\boldsymbol{v})=|\langle\boldsymbol{v},\boldsymbol{\varphi}^b_{(s,e]}\rangle|=C^b_{(s,e]}(\boldsymbol{v})$$

1.3 变点检测方法

本文的主要思想是将数据 $Y=(Y_1,Y_2,\cdots,Y_T)$ 的局部与全局处理相结合, 将统计量与二元分割相结合用以检测数据中的多变点, 首先在全局阶段, 随机绘制一些子区间$(s,e]$, 此区间的数据为子样本$(Y_{e+1},Y_{e+2}\cdots,Y_s)$, $1\leqslant s<e\leqslant T$, 其中 s、e 是整数。在局部阶段, 假设在每个子样本中最多只有一个变点, 计算每个子样本的统计量,取每个子样本中统计量的最大值, 并给定一个阈值 ζ, 若其统计量最大值超过给的定阈值, 就可以确定在此区间内存在一个疑似变点, 并将此位置以及统计量值保存下来, 剔除没有超过阈值的子样本。根据每次在局部阶段运行出来的结果, 进行最优的筛选就会得出最终变点位置及个数。

运用对比函数 $C^b_{(s,e]}(\boldsymbol{v})$作为检验统计量来检测变点, 首先在 $0,1,\cdots,T$ 上随机生成 M 个左开右闭的区间$(s,e]$,对于 M 的选择, 与数据量 T 有关, 一般是随着 T 的增大而增加,通常情况下, 在数据长度为 1 000 左右时,设 $M=1\ 000$。记 F^M_T 为这 M 个区间的集合, $(s,e]\in F^M_T$, 并给定一个阈值 $\zeta=K(2\log T)^{1/2}$, 通常取 $K=1,1.3$(见文献[20]), 设 $S=\varnothing$用以存储每次运行出的候选点, $G=\varnothing$用以存储候选变点所对应的统计量的值, 检测步骤如下:

Step 1 随机选取 F^M_T 中的区间$(s_m,e_m]$并计算其统计量 $C^b_{(s_m,e_m]}(\boldsymbol{v})$的值。

Step 2　若 $\zeta_T^* = \max_{s_m<b\le e_m} C^b_{(s_m,e_m]}(\boldsymbol{v}) > \zeta$，则可能存在一个潜在的变点 $b^* = \arg\max_{s_m<b\le e_m} C^b_{(s_m,e_m]}(\boldsymbol{v})$，$S = S \cup \{b^*\}$，$G = G \cup \{\zeta_T^*\}$。

Step 3　以 b^* 为分割点，将整个数据分为 $(0, b^*]$，$(b^*, T]$ 两个区间段，然后分别从 Step 1 开始运行，直到运行完 F_T^M 中所有的区间。

最后，运用强化型施瓦茨信息准则（Strengthened Schwarz Information Criterion，sSIC）对集合 S 集进行最优筛选，得出最终变点。

1.4　强化型施瓦茨信息准则

由 1.3 节可知，初始变点估计的数量 $\hat{q}$ 和位置 $\hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_{\hat{q}}$ 取决于选择的阈值 ζ，在接下来的变点筛选中，设 $\zeta_T \in G$，对于给定的数据，记 $\hat{q}(\zeta_T) = \hat{q}$，用 $T(\zeta_T) = \{\hat{\tau}_1(\zeta_T), \hat{\tau}_2(\zeta_T), \cdots, \hat{\tau}_{\hat{q}(\zeta_T)}(\zeta_T)\}$ 表示变点检测步骤中用阈值 ζ_T 估计的变点位置的集合。作函数 $\zeta_T \longmapsto T(\zeta_T)$，仅在不连续的点上改变其值，例如，对于任意 $i = 1, 2, \cdots, N-1$，若 $\zeta_T^{(1)} > \zeta_T^{(2)} > \cdots > \zeta_T^{(N-1)}$，则 $T(\zeta_T^{(i)}) \neq T(\zeta_T^{(i+1)})$，对于任意的 $\zeta_T \in [\zeta_T^{(i)}, \zeta_T^{(i+1)})$，有 $T(\zeta_T) = T(\zeta_T^{(i)})$。

然而，阈值 $\zeta_T^{(i)}$ 是未知的并且取决于数据，因此，在预先指定的阈值范围内直接使用 1.3 节的检测方法通常不会得到最优的结果。而且，从计算的角度来看，重复应用该方法来进行最优选择，并不能得到一个好的结果，因为希望 $\zeta_T^{(i)}$ 和 $\zeta_T^{(i+1)}$ 的解对于大多数 i 都是相似的。这些问题可以通过加强型施瓦茨准则来进行避免。

假设有已形成的选择方案 $T(\zeta^{(1)}), T(\zeta^{(2)}), \cdots, T(\zeta^{(N)})$，运用 sSIC 来最小化的 $T(\zeta_T^{(k)})$，定义如下，设 $\hat{q}_k = |T(\zeta_T^{(k)})|$，$k = 1, 2, \cdots, N$，估计的变点为 $\hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_{\hat{q}_k} \in T(\zeta_T^{(k)})$，对于一些预先给定的 $\alpha \ge 1$，$\hat{\tau}_0 = 0$，$\hat{\tau}_{\hat{q}_{k+1}} = T$，强化型施瓦茨信息准定义则为

$$\mathrm{sSIC}(k) = -2\sum_{j=1}^{\hat{q}_k+1} \log\{l(Y_{\hat{\tau}_{j-1}+1}, \cdots, Y_{\hat{\tau}_{j-1}+2}, Y_{\hat{\tau}_j}; \hat{\boldsymbol{\Theta}}_j)\} + n_k \log^{\alpha}(T)$$

其中，n_k 表示估计参数的总数，包括 $\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2, \cdots, \hat{\boldsymbol{\Theta}}_{\hat{q}_{k+1}}$ 中的变点和自由参数的位置，当 $\alpha = 1$ 时，就是所熟知的施瓦茨信息准则，选择最优变点就是使得 sSIC 达到最小时 k 的取值。

1.5　计算复杂度分析

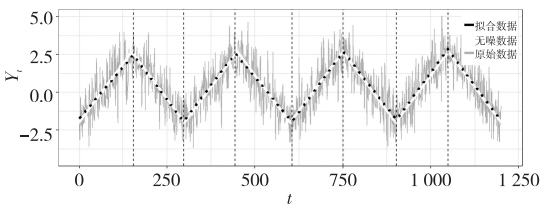
算法的计算复杂度描述的是该算法的运行时间，由于本文方法在计算时，是分别对每个区间的局部进行计算，所以计算复杂度相对较低。首先，需要对 M 个随机选取的左开右闭区间内的所有点进行分析，这些区间的起始点和结束点分别在 $\{0, 1, \cdots, T-1\}$ 和 $\{1, 2, \cdots, T\}$ 中，此计算复杂度取决于计算单个区间的对比函数的成本。在本文的研究中，该成本与区间长度呈线性关系，例如，计算单个 $C^b_{(s,e]}(\boldsymbol{v})$ 的成本是 $O(e-s)$。由于程序中随机选取的每个区间平均约有 $O(T)$ 个点，所以平均每个区间的计算复杂度为 $O(T)$，由于一个区间的计算完全独立于另一个区间的计算，所以该程序是以一种非常简单的“并行”方式来运行，又因为总共需要计算 M 个区间，因此，本文方法的计算复杂度是 $O(MT)$。

2　模拟研究

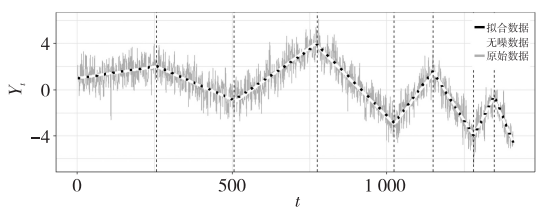
在本节中，根据模型式(1)，产生两组模拟数据以检测本文方法，如图 1 的两幅图所示，其中波动较大的灰色实线（有噪音）表示所模拟的原始数据 Y_t，中间的白色实线表示无噪音的分段线性数据 f_t，与白色实线相近的黑色虚线表示的是根据本文的方法，拟合出来的分段线性数据。从图 1 中可以明显看出，本文的方法对数据转折点（变点）处的识别很精准，使得运用线性回归来拟合两变点之间数据有着不错的效果，中间的白色实线与黑色虚线几乎重叠。

两组数据都是具有 7 个变点的分段线性数据，对于第一组，每两个变点间线性趋势的倾斜度都相同，也就是斜率绝对值相同，而第二组数据中，线性趋势段的倾斜度则不一定相同，设两组的误差 $\varepsilon_t \sim N(0, \sigma_t^2)$，且噪声参数 $\sigma_t = 1$，随机产生的区间数 $M = 1\,000$。对于第一组，模拟产生了 1 200 个数据，变点之间的距离都相同，设置其变点位置向量 $\boldsymbol{\tau} = (150, 300, 450, 600, 750, 900, 1\,050)$，每段斜率为 $|k| = 1/32$，初始截距为 -2，运用本文的方法，检

测得出估计的变点位置为 $\hat{\boldsymbol{\tau}}=(150,304,446,601,750,898,1\,050)$，如图1(a)所示，由此可得估计变点位置的最大误差为 $\max|\boldsymbol{\tau}_i-\hat{\boldsymbol{\tau}}_i|=3$。对于第二组，模拟产生1 048个数据，设其变点的位置向量为 $\boldsymbol{\tau}=(256,512,768,1\,024,1\,152,1\,280,1\,344)$，其分段斜率为 $k_i=(-1)^{i-1}(2i-1)/256$，$i=1,2,\cdots,8$，初始截距为1，检测结果得 $\hat{\boldsymbol{\tau}}=(256,513,765,1\,027,1\,149,1\,279,1\,344)$，如图1(b)所示，同样，可得估计变点位置的最大误差为 $\max|\boldsymbol{\tau}_i-\hat{\boldsymbol{\tau}}_i|=3$，由此可见该方法对于具有分段线性趋势的数据，检测结果非常良好。



(a)分段趋势相同(斜率绝对值相同)的变点检测



(b)分段趋势不同(斜率绝对值不同)的变点检测

图1 分段线性趋势模拟结果

Fig. 1 Simulation results of piecewise linear trend

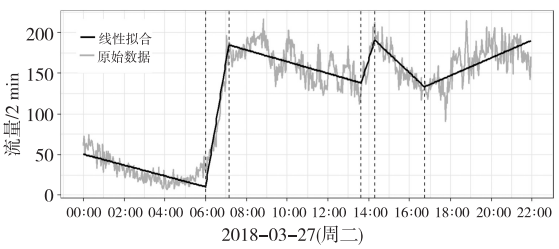
3 实例分析

城市道路的交通状态是交通运行系统中通行能力的体现，交通流变点就是由于某种变化而导致的，这种变化有可能是交通事故、自然灾害、交通管控等，有效及时地分析出交通流的突变情况对提升道路交通的通行能力有很大帮助。

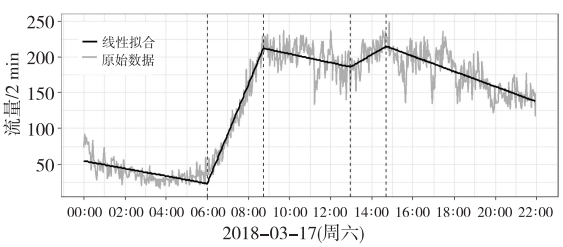
选取深圳市北环大道新洲立交的交通流卡口数据作为研究对象，以2018-03-17(周六)和2018-03-27日(周二)00:00—22:00的数据为例，每日共660个数据，对道路卡口每2 min的车流量进行变点检测(数据来源于2018年深圳杯竞赛D题)。

从图2可知，工作日和周六的数据分布以及变化在某些时段不太一致，在工作日(周二)，检测出的变点分别为06:00,07:08,13:38,14:24,16:44，对照图2(a)，可以得出在时间段06:00—07:08期间，车流量的增量非常明显，在13:38—14:24时间段，有短暂的增加，在16:44以后，道路车流量也增加地比较明显，可以得出，在工作日，该道路的早中晚高峰期比较明显，这与实际情况完全吻合。

在休息日(周六)，检测出的变点分别为06:00,08:44,12:58,14:42,相比工作日，变点减少了一个，而且明显可以看出，在06:00—08:44和12:58—14:42这时间段的车流量增长比工作日要缓慢许多，这是由于周末大家都没有上班，车流量的增长速度相对工作日就比较缓慢，而在下午却没有晚高峰，由于下午没有下班的高峰期，所以下午就没有变点的出现。



(a)2018年3月27日(周二)交通流变点检测



(b)2018年3月17日(周六)交通流变点检测

图2 深圳市北环大道新洲立交车流量变点检测

Fig. 2 Change point detection of traffic flow at Xinzhou Interchange in North Ring Avenue, Shenzhen

综上可得，在工作日和非工作日的交通流量，在增长速度方面的差异比较大，工作日的变化情况比较明显，该方法能很好地检测出交通流中的变化情况，如果出现了交通事故等问题，交通流量情况一定会在数据中体现出来，只需运用此方法就可知道交通中出现的事故等,并且出行人可以合理调整自己的出行时间，避开出行的高峰期，交管部门也可根据此实际情况对交通进行有效调控。

4　结束语

首先通过计算局部统计量 $C^{b}_{(s,e]}(\boldsymbol{v})$ 检测单变点,运用二元分割方法将其推广到全局的多变点检测,同时将阈值准则和 sSIC 相结合得出最终变点。将两种准则结合起来,得到了一种具有较好实际性能和最少参数选择的通用方法。由于本文的方法是将局部和全局相结合, 所以使得计算复杂度降为 $O(MT)$,因此可以运用于大数据集, 且对具有分段线性趋势的数据, 检测出的变点比较准确。然而本文只是针对一维数据进行讨论, 并没有涉及多维数据的变点问题, 所以可以将本文方法推广到多维数据变点的情况。

参考文献(References):

[1] PAGE E S. Continuous Inspection Schemes [J]. Biometrika, 1954, 41(1/2):100—115

[2] 胡尧,邓春霞,李丽. 非参数回归模型均值与方差双重变点的估计[J]. 应用概率统计, 2018, 34(3): 251—264
HU Y, DENG C X, LI L. Estimation of Change Point in Mean and Variance of Non-parametric Regression Mode [J]. Chinese J Appl Probab Statist, 2018, 34(3): 251—264 (in Chinese)

[3] 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: 20—30
TAN C C. Statistical Inference of Change Point Problem and Its Application in Finance[D]. Hefei: University of Science and Technology of China, 2007: 20—30 (in Chinese)

[4] JAMES N A, KEJARIWAL A, MATTESON D S. Leveraging Cloud Data to Mitigate User Experience from 'Breaking Bad' [C]//2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016: 3499—3508

[5] YU M, RUGGIERI E. Change Point Analysis of Global Temperature Records [J]. International Journal of Climatology, 2019, 39(8): 3679—3688

[6] BARDWELL L, FEARNHEAD P. Bayesian Detection of Abnormal Segments in Multiple Time Series [J]. Bayesian Analysis, 2017, 12(1): 193—218

[7] MATTESON D S, JAMES N A, NICHOLSON W B, et al. Locally Stationary Vector Processes and Adaptive Multivariate Modeling [C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 8722—8726

[8] JACKSON B, SCARGLE J D, BARNES D, et al. An Algorithm for Optimal Partitioning of Data on an Interval [J]. IEEE Signal Processing Letters, 2005, 12(2): 105—108

[9] KILLICK R, FEARNHEAD P, ECKLEY I A. Optimal Detection of Changepoints with a Linear Computational Cost[J]. Journal of the American Statistical Association, 2012, 107(500): 1590—1598

[10] MAIDSTONE R, HOCKING T, RIGAILL G, et al. On Optimal Multiple Changepoint Algorithms for Large Data [J]. Statistics and Computing, 2017, 27(2): 519—533

[11] BAI J, PERRON P. Estimating and Testing Linear Models with Multiple Structural Changes [J]. Econometrica, 1998(5): 47—78

[12] KIM S J, KOH K, BOYD S, et al. l_1 Trend Filtering[J]. SIAM Review, 2009, 51(2): 339—360

[13] TIBSHIRANI R J. Adaptive Piecewise Polynomial Estimation via Trend Filtering [J]. The Annals of Statistics, 2014, 42(1): 285—323

[14] FEARNHEAD P, MAIDSTONE R, LETCHFORD A. Detecting Changes in Slope with an L_0 Penalty [J]. Journal of Computational and Graphical Statistics, 2019, 28(2): 265—275

[15] SPIRITI S, EUBANK R, SMITH P W, et al. Knot Selection for Least-squares and Penalized Splines [J]. Journal of Statistical Computation and Simulation, 2013, 83(6): 1020—1036

[16] VOSTRIKOVA L Y. Detecting "Disorder" in Multidimensional Random Processes [C]// Doklady Akademii Nauk. Russian Academy of Sciences, 1981, 259(2): 270—274

[17] OLSHEN A B, VENKATRAMAN E S, LUCITO R, et al. Circular Binary Segmentation for the Analysis of

Array-based DNA Copy Number Data[J]. Biostatistics, 2004, 5(4): 557—572

[18] FRYZLEWICZ P. Wild Binary Segmentation for Multiple Change-point Detection[J]. The Annals of Statistics, 2014, 42(6): 2243—2281

[19] BARANOWSKI R, CHEN Y, FRYZLEWICZ P. Narrowest-over-threshold Detection of Multiple Change Points and Change-point-like Features[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2019, 81(3): 649—672

[20] HAMPEL F R. The Influence Curve and Its Role in Robust Estimation[J]. Journal of the American Statistical Association, 1974, 69(346): 383—393

Multiple Change-points Detection in Piecewise Linear Trends Based on Binary Segmentation

LIU Wei[1], HU Yao[1,2], HU Qian[1]

(1. School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China;
2. Guizhou Provincial Key Laboratory of Public Big Data, Guiyang 550025, China)

Abstract: Change point detection has always been one of the hot research topics in statistics. In actual data, there is usually a linear increase or decrease trend in a certain segment, the starting and ending point of this trend is unknown, thus, aiming at this kind of one-dimensional data with piecewise linear trend, a multiple change point detection method is proposed. Based on the statistics constructed by the generalized log-likelihood ratio, this method combines the binary segmentation method, threshold criterion, and Strengthened Schwarz information criterion to quickly and effectively detect multiple change points in the data. Numerical simulation results show that the method is very accurate in detecting the position and number of change points for the data with piecewise linear trends, and the detection results are satisfactory. Finally, by taking the traffic flow data of Xinzhou Interchange of North Ring Avenue in Shenzhen as an example, the distribution characteristics of the change point of the area on working days and non-working days are analyzed. The analysis results are consistent with the actual situation and can provide reference opinions for the relevant work of the traffic management departments.

Key words: multiple change-points test; strengthened Schwarz information criterion; binary segmentation; piecewise-linear trend subsection

责任编辑:罗姗姗

引用本文/Cite this paper:

刘伟,胡尧,胡倩. 基于二元分割检测分段线性趋势中的多变点[J]. 重庆工商大学学报(自然科学版),2020,37(6):32—38

LIU W, HU Y, HU Q. Multiple Change-points Detection in Piecewise Linear Trends Based on Binary Segmentation[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2020, 37(6): 32—38