Pure Mathematics 理论数学, 2018, 8(2), 136-148
Published Online March 2018 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2018.82018

Hans汉斯

Multiple Change-Points Detection of Piecewise Stationary Time Series

Nan Wu[1], Yao Hu[1,2*], Dan Wang[1]

[1]School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
[2]Guizhou Provincial Key Laboratory of Public Big Data, Guiyang Guizhou
Email: *yhul@gzu.edu.cn, 331914178@qq.com

Received: Feb. 23rd, 2018; accepted: Mar. 9th, 2018; published: Mar. 14th, 2018

Abstract

The change-point problem has a wide range of applications in the industrial, financial, meteorology and other fields. A method for estimating the numbers, locations of change-point by building the Likelihood ratio scan (LRS) statistics, combined with the Minimum description length (MDL) principle has been proposed. It reduces the computationally infeasible global multiple-change-point estimation problem to a number of single-change-point detection problems in various local windows by effective segmentation. In order to provide more information for describing change points, we have constructed confidence intervals for each of the change points. Finally, extensive simulation studies and example analysis of traffic show the LRS usability practice.

Keywords

Multiple Change-Points, Likelihood Ratio, Confidence Intervals, Minimum Description Length Criterion

分段平稳时间序列中的多变点检测

吴　楠[1]，胡　尧[1,2*]，王　丹[1]

[1]贵州大学数学与统计学院，贵州　贵阳
[2]贵州省公共大数据重点实验室，贵州　贵阳
Email: *yhul@gzu.edu.cn, 331914178@qq.com

收稿日期：2018年2月23日；录用日期：2018年3月9日；发布日期：2018年3月14日

*通讯作者。

文章引用: 吴楠, 胡尧, 王丹. 分段平稳时间序列中的多变点检测[J]. 理论数学, 2018, 8(2): 136-148.
DOI: 10.12677/pm.2018.82018

摘　要

变点问题在工业、金融、气象等领域有着广泛的应用。针对分段平稳时间序列的多变点检测，提出一种通过构建似然比扫描(LRS)统计量，结合最小描述长度(MDL)准则对变点数量、位置进行估计的方法，将计算上不可行的全局多变点估计问题通过有效分段降为各局部窗口中的多个单变点检测问题。同时对每个估计变点构建了置信区间，为描述变点提供更多信息。最后通过大量数值模拟和交通实例分析证明方法的有效性。

关键词

多变点，似然比，置信区间，最小描述长度准则

1. 引言

近年来，变点检测一直被认为是计量经济学，生物学和统计学的一个重要问题。大量文献探索了时间序列模型中变点的检测。变点估计的第一篇文章是由 Hinkley (1970)提出的，他研究 *i.i.d* 随机变量序列中变点的极大似然估计，并证明了估计变点依分布收敛于双侧随机游走的最大值，在正态假设下，表明其极限分布可以通过数值方法获得[1]。Hinkley (1970)使用类似的方法研究了服从二项分布的随机变量序列，并给出了极限分布的可计算形式[2]。但是，对于非正态或非二项分布的情况，其结果不能作为变点的统计推断。直到 Yao (1987)表明当参数变化幅度 d 很小时，Hinkley (1970)的极限分布可以作为变点估计分布的一个良好近似[3]。在时间序列领域，Picard (1985)首先研究了自回归(AR)模型中变点的极大似然估计。她假设参数变化幅度是 d_n，d_n 取决于样本量 n，当 $n\to\infty$ 时，$d_n\to 0$，结果得到与 Yao (1987)相同的极限分布，这为本文后期对变点构建置信区间提供理论支持[4]。Davis (2006)提出利用最小描述长度(MDL: Minimum description length)准则来检测非线性时间序列中的多变点[5]。Fryzlewicz (2014)提出 wild 二元分割(WBS: Wild Binary Segmentation)方法，省去复杂的优化过程，将非平稳时间序列随机分割为分段平稳序列，运用乘积性模型将时间序列自协方差结构中的变点检测问题转化为检测小波周期图中的变点问题[6]。Yau (2016)研究了非平稳时间序列中的多变点问题，通过似然比扫描方法将原始序列分解为一段一段的自回归过程，分别对各段 AR 过程进行建模，AR 模型参数发生改变的点即为变点。随着这些方法的产生，非平稳时间序列中的多变点研究问题得到了快速发展[7]。

实际上，仅仅检测出变点位置并不能提供完整的信息，相比之下，置信区间可以提供更多信息来对变点进行描述。为了获得置信区间，则需要得到变点估计的渐近分布。Hinkley (1970)、Picard (1985)和 Yao (1987)分别研究了估计变点的收敛情况及极限分布，指出变点估计依分布收敛于双侧随机游走的最大值，其极限分布参见 Yao (1987)。

本文首先介绍了变点模型及其渐近理论。其次，结合似然比方法构造似然比扫描(LRS: Likelihood ratio scan)统计量进行变点检测，并给出相应的渐近性质及似然比扫描方法的详细执行步骤：1) 使用似然比扫描统计量获得初步变点估计(可能存在过估计问题，将一些不是异常点识别为变点)；2) 采用 MDL 准则

进行模型选择过程，进而得到一组一致变点估计；3) 为每个估计的变化点构建置信区间。最后，利用大量数值模拟和交通实例验证 LRS 方法的有效性、优良性和实用性。

2. 模型介绍

2.1. 基本假设

假设时间序列 $\{X_t : t=1,2,\cdots\}$ 是 $\mathcal{F}_t$ 可测的，并且严格平稳的，遍历的，可表示为

$$X_t = g(\boldsymbol{\theta}, \boldsymbol{X}_t, \epsilon_t) \tag{1}$$

其中 $\mathcal{F}_t$ 是由 $\{\epsilon_t, \epsilon_{t-1}, \cdots\}$ 生成的 σ 域，$\boldsymbol{X}_t = (X_t, X_{t-1}, \cdots)$，$\boldsymbol{\theta}$ 是 $p\times 1$ 维的未知参数向量，$\{\epsilon_t\}$ 是独立同分布的。序列 $\{X_t\}$ 的结构由可测量的函数 g 和参数 $\boldsymbol{\theta}$ 刻画，假设参数空间 Θ 是 $\mathbb{R}^p$ 的有界紧集，且 g 是关于 $\boldsymbol{\theta}$ 连续的。当真实参数 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 时，用 $M(\boldsymbol{\theta}_0)$ 表示模型(1)，同理 $\boldsymbol{\theta}=\boldsymbol{\theta}_1$ 时，用 $M(\boldsymbol{\theta}_1)$ 表示模型(1)。

首先考虑分段平稳过程中的单变点问题。设 $\{X_1, \cdots, X_n\}$ 是由两个独立过程组成的随机样本，记 $\tau_1^0 \in \{1,2,\cdots,n-1\}$ 为真实变点，满足 $\delta n < \tau_1^0 < (1-\delta)n, \delta > 0$，则单变点问题模型为

$$\left\{X_1, \cdots, X_{\tau_1^0}\right\} \in M(\boldsymbol{\theta}_1^0), \quad \left\{X_{\tau_1^0+1}, \cdots, X_n\right\} \in M(\boldsymbol{\theta}_2^0) \tag{2}$$

其中 $\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \in \Theta$ 是两段序列的未知参数，且 $\boldsymbol{\theta}_1^0 \neq \boldsymbol{\theta}_2^0$。

记条件似然函数 $l_t(\boldsymbol{\theta}) \equiv \log f_{\boldsymbol{\theta}}(X_t \mid \boldsymbol{X}_{t-1})$，其中 $f_{\boldsymbol{\theta}}$ 是 X_t 给定过去观测值的条件密度函数。则单变点模型(2)的似然函数为

$$L_n(\tau, \theta_1, \theta_2) = L_{1n}(\tau, \theta_1) + L_{2n}(\tau, \theta_2)$$

式中 $L_{1n}(\tau, \boldsymbol{\theta}_1) = \sum_{t=1}^{\tau} l_t(\boldsymbol{\theta}_1)$，$L_{2n}(\tau, \boldsymbol{\theta}_2) = \sum_{t=\tau+1}^{n} l_t(\boldsymbol{\theta}_2)$ 分别是两段序列的条件对数似然函数。对于给定的 τ，$\hat{\boldsymbol{\theta}}_{1n}(\tau) = \operatorname{argmax}_{\boldsymbol{\theta}_1} L_{1n}(\tau, \boldsymbol{\theta}_1)$，同理 $\hat{\boldsymbol{\theta}}_{2n}(\tau) = \operatorname{argmax}_{\boldsymbol{\theta}_2} L_{2n}(\tau, \boldsymbol{\theta}_2)$，则变点 τ 可由(3)式进行估计

$$\hat{\tau} = \arg\max_{1\le\tau\le n} L_n\left[\tau, \hat{\boldsymbol{\theta}}_{1n}(\tau), \hat{\boldsymbol{\theta}}_{2n}(\tau)\right] \tag{3}$$

变点前后两段的参数估计分别为 $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}_{1n}(\tau)$，$\hat{\boldsymbol{\theta}}_2 = \hat{\boldsymbol{\theta}}_{2n}(\tau)$。$\hat{\tau}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ 的一致性证明参见 Ling (2016) [8]。

2.2. 变点估计的渐近分布

对于单变点模型(2)，定义随机游走

$$W_\tau = \begin{cases} \sum_{t=1}^{\tau}\left(l_t(\boldsymbol{\theta}_1^0) - l_t(\boldsymbol{\theta}_2^0)\right), & \tau > 0, \\ 0, & \tau = 0, \\ \sum_{t=\tau}^{-1}\left(l_t(\boldsymbol{\theta}_2^0) - l_t(\boldsymbol{\theta}_1^0)\right), & \tau < 0. \end{cases} \tag{4}$$

当 $\tau > 0$ 时 $X_t \in M(\boldsymbol{\theta}_2^0)$，$\tau < 0$ 时 $X_t \in M(\boldsymbol{\theta}_1^0)$。参照 Ling (2016)的定理 2.2(b)，对于固定的 $\boldsymbol{\theta}_1^0$，$\boldsymbol{\theta}_2^0$ 有

$$\hat{\tau}_1 - \tau_1^0 \xrightarrow{d} \arg\max_{\tau\in\mathbb{Z}} W_\tau \tag{5}$$

注意到 W_τ 的极限分布依赖于未知参数 $\boldsymbol{\theta}_1^0$，$\boldsymbol{\theta}_2^0$，并且不能得出分布函数的解析式，因此，实践中难以直接使用(5)构造置信区间。Ling (2016)定理 3.1 推导出了当参数变化很小时 W_τ 的近似分布。具体而言，如果 $\boldsymbol{\theta}_1^0 - \boldsymbol{\theta}_2^0 = O\left(1/\sqrt{n}\right)$，那么对于任意给定的 $M\ (M>0)$

$$\left(\hat{d}'\hat{\Sigma}\hat{d}'\right)^2\left(\hat{d}'\hat{\Omega}\hat{d}'\right)^{-1}\left(\arg\max_{r\in[-M,M]} W_{\lfloor \hat{m}r \rfloor}\right) \xrightarrow{d} \arg\max_{r\in[-M,M]}\left\{B(r) - \frac{1}{2}|r|\right\} \tag{6}$$

其中 $\hat{d}=\hat{\theta}_1-\hat{\theta}_2$，$\hat{\Sigma}=\frac{1}{2M}\sum_{t=-M}^{M}\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big|_{\hat{\theta}_2}$，$\hat{\Omega}=\frac{1}{2M}\sum_{t=-M}^{M}\left(D_t\left(\hat{\theta}_2\right)-\bar{D}\right)\left(D_t\left(\hat{\theta}_2\right)-\bar{D}\right)'$，$D_t(\theta)=\frac{\partial l_t(\theta)}{\partial\theta}$，$\bar{D}=\sum_{-M}^{M}\frac{D_t\left(\hat{\theta}_2\right)}{2M}$，$B(r)$ 是 $\mathbb{R}$ 上的标准布朗运动，$\hat{m}=\left(\hat{d}'\hat{\Sigma}\hat{d}'\right)^{-2}\left(\hat{d}'\hat{\Omega}\hat{d}'\right)$ (变点数目)。

进一步，记 $\arg\max_\tau W_\tau$ 的分布为 $F_d(x)$，则当 $s>M$ 时，有

$$\left|F_d(x)-P\left(\arg\max_{\tau\in[-s,s]}W_{\lfloor\tau\rfloor}\le x\right)\right|\le\varepsilon$$

因此

$$\begin{aligned}&\left|F_d(x)-P\left(m\arg\max_{r\in[-M,M]}W_{\lfloor\hat{m}r\rfloor}x\right)\right|\\&=\left|F_d(x)-P\left(\arg\max_{r\in[-mM,mM]}W_{\lfloor r\rfloor}\le x\right)\right|\le\varepsilon\end{aligned}$$

当 M 充分大时，$r\notin[-M,M]$ 的概率很小，所以当 d 较小时，

$$F_d(x)\approx P\left(\arg\max_{r\in R}\left[B(r)-|r|/2\right]\le x\right)$$

成立。Yao (1987)证明了 $\arg\max_{r\in R}\left[B(r)-|r|/2\right]$ 的分布 $F(x)$ 具有密度函数：

$$f(x)=\frac{3}{2}e^{|x|}\Phi\left(\frac{3}{2}\sqrt{|x|}\right)-\frac{1}{2}\Phi\left(\frac{\sqrt{|x|}}{2}\right)\tag{7}$$

其中 $\Phi(x)=\int_x^\infty\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)du$，$x\in R\equiv(-\infty,\infty)$。故当 M 充分大，d 较小时，可以用 $F(x)$ 来近似 $\left(\hat{d}'\hat{\Sigma}\hat{d}'\right)^2\left(\hat{d}'\hat{\Omega}\hat{d}'\right)^{-1}\left(\hat{\tau}_1-\tau_1^0\right)$ 的分布。利用 $F(x)$ 对变点估计进行置信区间构造，即 $(1-\alpha)\%$ 置信度的置信区间为

$$CI=\left[\hat{\tau}_1-\left[\Delta F_{\alpha/2}\right]-1,\hat{\tau}_1+\left[\Delta F_{\alpha/2}\right]+1\right]\tag{8}$$

其中 $\Delta=\left(\hat{d}'\hat{\Omega}\hat{d}'\right)\left(\hat{d}'\hat{\Sigma}\hat{d}'\right)^{-2}$，$F_{\alpha/2}(x)$ 是 $\arg\max_{r\in\mathbb{R}}\left[B(r)-|r|/2\right]$ 的 $\frac{\alpha}{2}$ 分位数。

3. 基于似然比扫描 (LRS)方法的多变点估计

结合 AR 过程的计算优势和强稳健性，考虑将原始序列通过合理分割变为分段平稳 AR 过程，即将原始序列的多变点估计变为局部单变点估计。

3.1. 基本假设

假设观测值 $\{X_t\}_{t=1,\cdots,n}$ 可以划分为 $m+1$ 个平稳过程。第 j 段序列观测值 $\boldsymbol{X}_j=\left\{X_{\tau_{j-1}+1},\cdots,X_{\tau_j}\right\}\in M\left(\theta_j\right)$，其中 $j=1,\cdots,m$，τ_j 是第 j 个变点，即在 τ_j 处第 j 段 AR 过程 $\mathrm{AR}\left(p_j\right)$ 发生突变，变为第 $j+1$ 段 AR 过程 $\mathrm{AR}\left(p_{j+1}\right)$。令 $\tau_0\triangleq0$，$\tau_{m+1}\triangleq n$。记 $\mathcal{J}=(\tau_1,\cdots,\tau_m)$ 为变点的集合。定义 λ_j 为第 j 个变点的相对位置，使得 $\tau_j=\lfloor\lambda_j n\rfloor$，$j=0,\cdots,m+1$，且满足当 $\epsilon_\lambda>0$ 时 $\min_{j=0,\cdots,m}\left(\lambda_{j+1}-\lambda_j\right)>\epsilon_\lambda$。

3.2. 基于似然比扫描统计量的多变点检测步骤

本节，主要介绍了 LRS 方法检测多变点的三个步骤，其变点估计及渐近性质将在下节讨论。

第一步：使用似然比扫描统计量获得所有变点集合

分别定义扫描窗口及其相应的观测值为

$$W_t(h)=\{t-h+1,\cdots,t+h\},\quad X_{W_t(h)}=\{X_{t-h+1},\cdots,X_{t+h}\}$$

其中 $t=h,\cdots,n-h$，h 称为窗口半径，$h=h(n)$ 依赖于样本量 n。定义扫描窗口 $W_t(h)$ 的似然比扫描统计量

$$S_h(t)=\frac{1}{h}L_{1h}\left(t,\hat{\boldsymbol{\theta}}_1\right)+\frac{1}{h}L_{2h}\left(t,\hat{\boldsymbol{\theta}}_2\right)-\frac{1}{h}L_{\cdot h}\left(t,\hat{\boldsymbol{\theta}}\right)$$

$L_{1h}\left(t,\hat{\boldsymbol{\theta}}_1\right)$，$L_{2h}\left(t,\hat{\boldsymbol{\theta}}_2\right)$，$L_{\cdot h}\left(t,\hat{\boldsymbol{\theta}}\right)$ 分别是相应观测序列 $\{X_s\}_{t-h+1,\cdots,t}$，$\{X_s\}_{t+1,\cdots,t+h}$，$\{X_s\}_{W_t(h)}$ 的条件似然函数。具体地说，样本 $z=\{z_1,\cdots,z_n\}$ 的条件似然函数为

$$\boldsymbol{L}(\boldsymbol{\theta})=\sum_{t=1}^{n}l_t(\boldsymbol{\theta})\equiv\sum_{t=1}^{n}\log f_{\boldsymbol{\theta}}\left(z_t\mid z_{t-1},z_{t-2},\cdots\right)$$

其中 $f_{\boldsymbol{\theta}}\left(z_t\mid z_{t-1},z_{t-2},\cdots\right)$ 是给定过去观测值的条件密度函数，且当 $s\le 0$ 时 $z_s=0$。

接下来，使用扫描统计量 $S_h(t)$ 对序列进行扫描，从而获得一组似然比扫描统计量值的序列 $\left(S_h(h),S_h(h+1),\cdots,S_h(n-h)\right)$，如果 t 是变点，$S_h(t)$ 往往会较大。若 $2h<n\epsilon_\lambda$，则每个扫描窗口内至多存在一个变点。因此 $S_h(\cdot)$ 的局部最大值构成一组变点估计，记局部变点估计如下

$$\hat{\mathcal{J}}^{(1)}=\left\{m\in\{h,h+1,\cdots,n-h\}:S_h(m)=\max_{t\in(m-h,m+h)}S_h(t)\right\}$$

当 $t\notin\left[h,\cdots,(n-h)\right]$ 时 $S_h(t)\triangleq 0$。即如果 $S_h(m)$ 是扫描窗口 $[m-h+1,m+h]$ 内的最大值，则 m 就是该扫描窗口的局部变点估计。记 $\hat{m}$ 为估计变点的数目，表示 $\hat{\mathcal{J}}^{(1)}$ 中的元素个数。

第二步：通过模型选择得到变点一致估计

寻找 m，$\mathcal{J}$ 的最佳集合即为模型选择问题。从上一步获得的局部变点估计集合 $\hat{\mathcal{J}}^{(1)}$ 通常会存在过估计问题，即将一些正常值也识别为变点，使得 $\hat{\mathcal{J}}^{(1)}$ 内不仅包含所有真实变点集合，还包含一些非变点。为了更准确地检测到真实变点，进一步利用合适的信息准则从 $\hat{\mathcal{J}}^{(1)}$ 中选取最佳的子集作为变点估计。最小描述长度(MDL)准则已经在很多实证研究中体现出显著的优势(如 Davis *et al.* (2006, 2008))，本文选取 MDL 准则进行模型选择过程。给定一组变点估计 $\mathcal{J}=(\tau_1,\cdots,\tau_m)$，MDL 准则定义为

$$\mathrm{MDL}(m,\mathcal{J})=\log(m)+(m+1)\log(n)+\sum_{j=1}^{m+1}\sum_{k=1}^{c_j}\log\left(\zeta_{j,k}\right)+\sum_{j=1}^{m+1}\frac{d_j}{2}\log(n_j)-\sum_{j=1}^{m+1}L_j\left(\hat{\boldsymbol{\theta}}_j;\boldsymbol{X}_j\right)$$

式中 $L_j\left(\hat{\boldsymbol{\theta}}_j;\boldsymbol{X}_j\right)$ 是第 j 段的似然函数，$n_1,\cdots,n_j$ 是每段分割的长度，d_j 是 $\boldsymbol{\theta}_j$ 的维数，$\zeta_{j,1},\cdots,\zeta_{j,c_j}$ 是整数值参数，分别确定第 j 段分割的模型参数[5] [9]。考虑本文将各段分为 AR 过程，整数值参数只有 AR 过程阶数 p_j；又 $\boldsymbol{\theta}_j=\left(\phi_{j,0},\phi_{j,1},\cdots,\phi_{j,p},\sigma_\epsilon^2\right)$，故 $d_j=p_j+2$。所以针对 AR 过程，其 MDL 表达式为

$$\mathrm{MDL}(m,\mathcal{J})=\log(m)+(m+1)\log(n)+\sum_{j=1}^{m+1}\log\left(p_j\right)+\sum_{j=1}^{m+1}\frac{p_j+2}{2}\log(n_j)-\sum_{j=1}^{m+1}L_j\left(\hat{\boldsymbol{\theta}}_j;\boldsymbol{X}_j\right)$$

给定局部变点估计 $\hat{\mathcal{J}}^{(1)}$，可以通过下式精确估计变点

$$\left(\hat{m}^{(2)},\hat{\mathcal{J}}^{(2)}\right)=\underset{m=|\mathcal{J}|,\mathcal{J}\subseteq\hat{\mathcal{J}}^{(1)}}{\operatorname{argmin}}\ \mathrm{MDL}(m,\mathcal{J})$$

因为 $\hat{\mathcal{J}}^{(1)}$ 所包含的因素已经远远少于样本容量，所以在 $\hat{\mathcal{J}}^{(1)}$ 上优化 MDL 使计算的复杂度大大降低，提高计算效率。本文采用 Yao (1984)和 Jackson (2005)提到的最优分割(OP)算法对 MDL 进行优化[10] [11]。

第三步：最终估计和置信区间构造

定义新扩展的局部窗口 $E_j(h)$ 和第 j 个变点估计 $\hat{\tau}_j^{(2)}\in\hat{\mathcal{J}}^{(2)}$ 的相应观测值 $X_{E_j(h)}$

$$E_j(h)=\left\{\hat{\tau}_j^{(2)}-2h+1,\cdots,\hat{\tau}_j^{(2)}+2h\right\},\quad X_{E_j(h)}=\left\{X_{\hat{\tau}_j^{(2)}-2h+1},\cdots,X_{\hat{\tau}_j^{(2)}+2h}\right\}$$

这保证了每个真实变点在相应扩展局部窗口 $E_j(h)$ 的 $\left(\frac{1}{4},\frac{3}{4}\right)$ 内的概率接近于 1。设 $L_j(\tau,\theta_1,\theta_2)=\sum_{t=\hat{\tau}_j^{(2)}-2h+1}^{\tau}l_t(\theta_1)+\sum_{t=\tau+1}^{\hat{\tau}_j^{(2)}+2h}l_t(\theta_2)$，定义最终变点估计为

$$\hat{\tau}_j^{(3)}=\underset{\tau\in\left[\hat{\tau}_j^{(2)}-h,\hat{\tau}_j^{(2)}+h\right]}{\arg\max}L_j\left(\tau,\hat{\boldsymbol{\theta}}_j,\hat{\boldsymbol{\theta}}_{j+1}\right)$$

其中 $\hat{\boldsymbol{\theta}}_j=\hat{\boldsymbol{\theta}}_j(\tau)=\operatorname{argmax}_{\theta_1}\sum_{t=\hat{\tau}_j^{(2)}-2h+1}^{\tau}l_t(\theta_1)$，$\hat{\boldsymbol{\theta}}_{j+1}$ 类似定义。此时，可利用 2.1 中的结论来获得每个最终的变点估计 $\hat{\tau}_j^{(3)}$ 的置信区间。

在最初扫描步骤中，$W_j(h)$ 的大小为 $2h$，对给定 h，每个时刻 t 的 $S_h(t)$ 的计算复杂度为 $O(h)$。在第二步中，使用 Jackson (2005)最优分割算法优化 MDL 时，最小化 MDL 需要 $O\left(\left(\hat{m}^{(1)}\right)^2 n\right)$ 的计算复杂度。在第三步中，由于计算被限制在扩展的局部窗口上，所以计算复杂度为 $O\left(\hat{m}^{(2)}h^2\right)$。综上，由于 $\hat{m}^{(1)}$ 和 $\hat{m}^{(2)}$ 都是有限的，因此整个检测到最终变点估计过程的总计算复杂度为 $O\left(nh+h^2\right)$。当 h 较小时，例如 $h=O\{\log(n)\}$，则完整的三步 LRS 方法需要使用动态规划算法(最优分割算法)的计算复杂度为 $O\{n\log(n)\}$，明显低于 $O\{n^2\}$ 的数量级。如 3.3 节所示，窗口半径 h 作为调整参数，其值选取对定理 3.1 是至关重要的，其中 d 未知的。但如果 h 的阶数大于 $O\{\log(n)\}$，例如 $h=O\left(d_2\log(n)^2\right)$，那 d_2 的取值不会影响到 LRS 方法的一致性。因此，随着样本量的增加，这个 h 的选择在理论上是合理的。经过大量模拟及实证研究发现通常 $d=2$ 时，各种模型和样本有较好的结果，因此建议当 $n>800$ 时使用 $\max\left\{50,2\log(n)^2\right\}$，当 $n\le 800$ 时使用 $\max\left\{25,2\log(n)^2\right\}$ 作为 h 的经验选择。

3.3. 渐近性质

本节主要讨论似然比扫描方法的渐近性质，给出相应定理以保证变点估计数目、位置、置信区间的一致性。

假设 3.1. 对任意两个连续的分段 $\boldsymbol{X}_j=\left\{X_{\tau_{j-1}+1},\cdots,X_{\tau_j}\right\}$，$\boldsymbol{X}_{j+1}=\left\{X_{\tau_j+1},\cdots,X_{\tau_{j+1}}\right\}$，当 $k\in\left\{\tau_{j-1}+1,\cdots,\tau_j\right\}$，条件似然函数的期望 $E\left[l_k(\boldsymbol{\theta})\right]$ 在 $\boldsymbol{\theta}=\boldsymbol{\theta}_j^0$ 处取得唯一的极大值，同理当 $k\in\left\{\tau_j+1,\cdots,\tau_{j+1}\right\}$，$E\left[l_k(\boldsymbol{\theta})\right]$ 在 $\boldsymbol{\theta}=\boldsymbol{\theta}_{j+1}^0$ 处取得唯一的极大值，且 $\boldsymbol{\theta}_j^0\ne\boldsymbol{\theta}_{j+1}^0$。此外有

$$\begin{cases}E\left[l_k\left(\boldsymbol{\theta}_{j+1}^0\right)\right]<E\left[l_k\left(\boldsymbol{\theta}_j^0\right)\right] & k\in\left\{\tau_{j-1}+1,\cdots,\tau_j\right\}\\ E\left[l_k\left(\boldsymbol{\theta}_{j+1}^0\right)\right]>E\left[l_k\left(\boldsymbol{\theta}_j^0\right)\right] & k\in\left\{\tau_j+1,\cdots,\tau_{j+1}\right\}\end{cases}$$

假设 3.2. 在任意一个分段中，$l_k(\boldsymbol{\theta})$ 是关于 $\{X_t\}$ 的连续可测函数，且关于 $\boldsymbol{\theta}$ 几乎处处二阶可微。

假设 3.3. 令 $Y_k(\boldsymbol{\theta})=l_k(\boldsymbol{\theta})-E\left[l_k(\boldsymbol{\theta})\right]$。对任意 $\boldsymbol{\theta}\in\Theta$，存在 $K>0$ 使得 $E\left(\mathrm{e}^{|Y_k(\boldsymbol{\theta})|}\right)\le K, k\in\mathbb{N}$ 成立。

假设 3.4. 对任意 $\boldsymbol{\theta}_j\in\Theta_j$，均存在可积函数 $G(\boldsymbol{X}_t)$，使得 $E\left(G(\boldsymbol{X}_t)\right)<\infty$，$\left|l_t(\boldsymbol{\theta}_j)\right|\le G(\boldsymbol{X}_t)$。

下面定理 3.1 确保了所有变点都可以在 $\hat{\mathcal{J}}^{(1)}$ 的 h 邻域中确定。

定理 3.1. 记真实变点集合为 $\mathcal{J}_0=\left(\tau_1^0,\cdots,\tau_{m_0}^0\right)$，通过第一步扫描统计量得到的局部变点集合为 $\hat{\mathcal{J}}^{(1)}=\left(\hat{\tau}_1^{(1)},\cdots,\hat{\tau}_{\hat{m}^{(1)}}^{(1)}\right)$，其中 $\hat{m}^{(1)}=\left|\hat{\mathcal{J}}^{(1)}\right|$。若假设 3.1~3.4 成立，且 $2h<n\epsilon_\lambda\left(\epsilon_\lambda>0\right)$，则存在 $d>0$，当 $h\ge d\log(n)$ 时有

$$\mathrm{P}\left(\underset{\tau\in\mathcal{J}_0,k=1,\cdots,\hat{m}^{(1)}}{\operatorname{maxmin}}\left|\tau-\hat{\tau}_k^{(1)}\right|<h\right)\to 1$$

由于变点之间的最小距离为 $n\epsilon_\lambda = O(n)$，所以真实变点数目 m_0 是有限的。但此时并不能保证 $\hat{m}^{(1)}$ 等于 m_0。也就是说，变点的数量可能被高估，存在过估计问题。接下来定理 3.2 阐述了基于 MDL 准则模型选择方法产生的变点数量和位置的一致性。

定理 3.2. 定理 3.1 成立条件下，当 $\epsilon_\lambda > 0$ 时，有 $\hat{m}^{(2)} \xrightarrow{p} m_0$。此外，若 $\hat{m}^{(2)} = m_0$，则有

$$\mathrm{P}\left(\max_{j=1,\cdots,m_0} \left| \hat{\tau}^{(2)} - \tau_j^0 \right| < h \right) \to 1$$

由于 $h \geq d\log(n)$，定理 3.2 意味着 $\max_{j=1,\cdots,m_0} \left| \hat{\tau}_j^{(2)} - \tau_j^0 \right| = O_p(h)$，与经典收敛速率 $O_p(1)$ 相比，显然是不理想的。不过尽管如此，区间 $\left[\hat{\tau}_j - h + 1, \hat{\tau}_j + h \right]$ 覆盖真实变点 τ_j^0 的概率接近依然 1，这使扩展的局部窗口产生变点一致最终估计和置信区间变得可行。

定理 3.3. 定理 3.2 成立条件下，若 $3h < n\epsilon_\lambda$，同时假设 3.4 成立，则有

$$\hat{\tau}_j^{(3)} - \tau_j^0 \xrightarrow{d} \arg\max_{\tau \in \mathbb{Z}} W_{j,\tau}$$

其中 $W_{j,\tau}$ 是一个随机游走如下所示

$$W_\tau = \begin{cases} \sum_{t=\tau_j^0+1}^{\tau_j^0+\tau} \left(l_t\left(\theta_j^0\right) - l_t\left(\theta_{j+1}^0\right) \right), & \tau > 0, \\ 0, & \tau = 0, \\ \sum_{t=\tau_j^0+\tau+1}^{\tau_j^0} \left(l_t\left(\theta_{j+1}^0\right) - l_t\left(\theta_j^0\right) \right), & \tau < 0. \end{cases}$$

特别地，此时 $\hat{\tau}_j^{(3)} - \tau_j^0 = O_p(1)$。

由于变点之间的最小距离远大于窗口半径 h，即 $n\epsilon_\lambda / h \to \infty$，扩展的局部窗口 $E_j(h)$ 之间的距离亦趋于无穷，所以在一些弱相依条件下，构造的 CI 是渐近独立的。此时可将 2.2 中获得单变点置信区间的方法直接应用至多变点模型(其实质也是局部单变点问题)。

4. 数值模拟

下面用数值模拟说明 LRS 方法的有效性。

4.1. 模型表达

模型 A：没有变点的平稳 AR(1)过程

模型 A 用来测试评估统计量在没有变点时的检测性能，即当序列无变点时，统计量是否能准确识别。AR(1)过程 $X_t = 0.4X_{t-1} + \varepsilon_t$，设样本量 $n = 1024$。

模型 B：分段平稳 AR(1)过程

$$X_t = \begin{cases} 0.4X_{t-1} + \varepsilon_t, & 1 \leq t \leq 400 \\ -0.6X_{t-1} + \varepsilon_t, & 401 \leq t \leq 612 \\ 0.5X_{t-1} + \varepsilon_t, & 613 \leq t \leq 1024 \end{cases}$$

模型 C：分段平稳 AR(2)过程

$$X_t = \begin{cases} 0.9X_{t-1} + \varepsilon_t, & 1 \leq t \leq 512 \\ 1.69X_{t-1} - 0.81X_{t-2} + \varepsilon_t, & 512 \leq t \leq 768 \\ 1.32X_{t-1} - 0.81X_{t-2} + \varepsilon_t, & 769 \leq t \leq 1024 \end{cases}$$

模型 D：分段平稳分段平稳 AR 过程(3 变点)

$$X_t=\begin{cases}1.399X_{t-1}-0.4X_{t-2}+\varepsilon_t, & 1\le t\le 125\\ 0.3X_{t-1}+0.3X_{t-2}+\varepsilon_t, & 126\le t\le 532\\ 0.9X_{t-1}+\varepsilon_t, & 533\le t\le 704\\ 0.1X_{t-1}-0.5X_{t-2}+\varepsilon_t, & 705\le t\le 1024\end{cases}$$

模型 E：分段平稳 ARMA(1，1)过程(2 变点)

$$X_t=\begin{cases}-0.9X_{t-1}+\varepsilon_t+0.7\varepsilon_{t-1}, & 1\le t\le 512\\ 0.9X_{t-1}+\varepsilon_t, & 513\le t\le 768\\ \varepsilon_t-0.7\varepsilon_{t-1}, & 769\le t\le 1024\end{cases}$$

在模型对比中加入 ARMA 过程，试图拓展统计量在 ARMA 过程中的变点识别应用问题。

上述各模型的变点检测模拟结果见表 1、图 1。

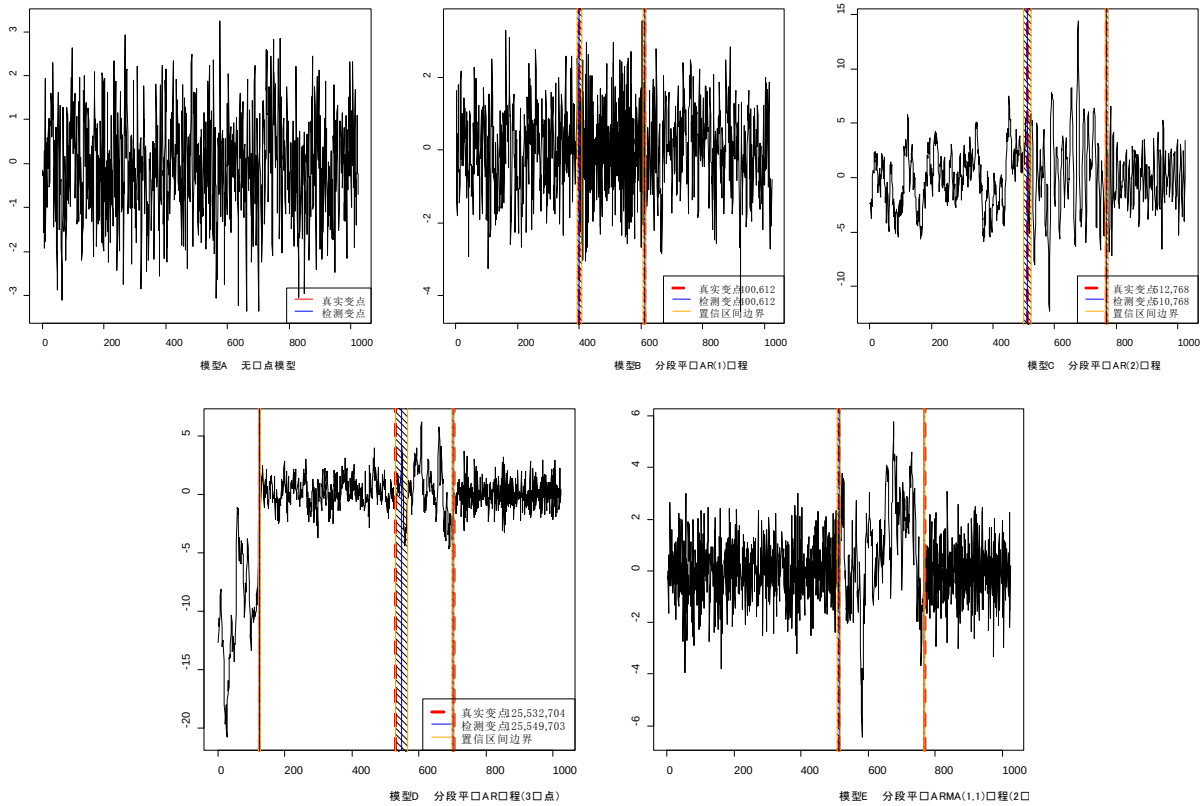

Figure 1. Location of true change-point and LRS estimation
图 1. 各模型真实变点与 LRS 方法估计变点位置

Table 1. Location of true change-point and LRS estimation
表 1. 各模型真实变点与 LRS 方法估计变点位置

模型	真实变点位置	估计变点位置	置信区间	模型	真实变点位置	估计变点位置	置信区间
模型 A	无	无	无	模型 D	125	125	[122, 128]
模型 B	400	400	[392, 408]		532	549	[531, 567]
	612	612	[607, 617]		704	703	[700, 706]
模型 C	512	510	[499, 525]	模型 E	512	511	[506, 516]
	768	768	[763, 773]		768	768	[766, 770]

结合表 1 和图 1 可看出，LRS 对于非平稳时间序列中的变点检测问题是有效的，不论是单变点、多变点还是没有变点的情形，均可以较准确的进行检测。在 ARMA 过程中的变点位置检测会出现偏移，但变点个数均正确，尽管准确率有所下降，不难看出检测到的变点仍在构造的置信区间内。

4.2. 置信区间

本节主要检查变点置信区间的覆盖准确性。通过上述各个模型生成数据。应用 LRS 方法，在每个估计变点周围利用定理 3.3 和 2.2 节的结论构建 90%置信区间。设置样本量 $n=1024$，分别各进行 100 次模拟。结果如表 2 所示。

表 2 展示了具体结论，表明最终变点估计 $\hat{\tau}_j^{(3)}$ 对 τ_0 做出了较好的估计，且置信区间覆盖概率较高，效果良好。

4.3. 模拟对比

定义利用渐近分布构造的置信区间覆盖率为 CR

$$CR=\frac{\#\text{置信区间覆盖真实变点}}{\text{模拟次数}}$$

趋近于 1，则认为估计效果比较好。

为对比检验 LRS 方法检测变点的准确率，依旧通过上述 5 个模型，同时利用 LRS 方法与经典 WBS 方法进行比较。各模型分别模拟 100 次，整理得到结果见表 3。

Table 2. Confidence interval coverage of each model
表 2. 各模型置信区间覆盖情况

模型	真实变点位置	变点估计均值	90%估计变点范围	90%置信区间均值	覆盖率	计算时间(s)
模型 A	无	无	无	无	无	无
模型 B	400	400.47	[387, 413]	[393.55, 407.39]	92%	387.96
	612	612.11	[599, 625]	[605.58, 618.64]	89%	
模型 C	512	512.26	[503, 520]	[501.11, 523.41]	92%	467.63
	768	768.61	[755, 782]	[760.92, 776.3]	93%	
模型 D	125	124.87	[116, 134]	[120.15, 129.59]	95%	563.92
	532	530.4	[493, 557]	[512.22, 548.58]	74%	
	704	704.74	[695, 714]	[700.16, 708.12]	89%	
模型 E	512	511.73	[504, 519]	[507.35, 516.09]	91%	491.58
	768	767.74	[761, 774]	[764.84, 770.62]	84%	

Table 3. Simulation result of LRS and WBS
表 3. LRS、WBS 模拟结果

变点个数	模型 A 无变点		模型 B 2 变点		模型 C 2 变点		模型 D 3 变点		模型 E 2 变点	
	LRS	WBS	LRS	WBS	LRS	WBS	LRS	WBS	LRS	WBS
0	100	55	0	0	0	0	0	0	0	0
1	0	12	0	0	0	0	0	0	0	14
2	0	13	100	56	100	86	12	17	100	53
3	0	9	0	28	0	13	88	67	0	26
≥4	0	11	0	16	0	1	0	16	0	7
CR	100%	51%	89%	50%	92%	72%	74%	60%	84%	41%

结果显示，当非平稳序列为分段平稳 AR 过程时 LRS 方法和 WBS 方法均能较好的检测出变点且位置估计效果都比较好，从变点个数检测和位置估计的准确率上来说，LRS 方法准确率明显高于 WBS 方法，WBS 方法存在较严重的变点数目高估问题，将一些本不是变点的点作为变点。通过上述典型模型的变点检测情况，明确了 LRS 方法的优良性，对各模型均能灵敏的对变点数目进行检测，给出了良好的变点位置估计，同时构建的置信区间覆盖率也较高，证明了 LRS 方法的实用性。

5. 交通流数据应用实例

以贵阳市宝山北路与东新区路交叉口南北方向车流量数据为例。选取 2016 年 4 月 4 日至 2016 年 4 月 10 日一周(星期一至星期日)车流量数据进行变点检测，验证 LRS 方法的实用性。数据为每天 00:00~23:55 每两分钟(共 720 个)过车数量。宝山北路与东新区路交叉口南北方向一周车流量时序图如图 2 所示。

考虑工作日、休息日车流量分布情况，分别选取星期四、五、六作为代表，同时利用 LRS 方法和 WBS 方法对车流量数据进行变点检测，分别得到各天的变点估计情况如图 3~5 所示。

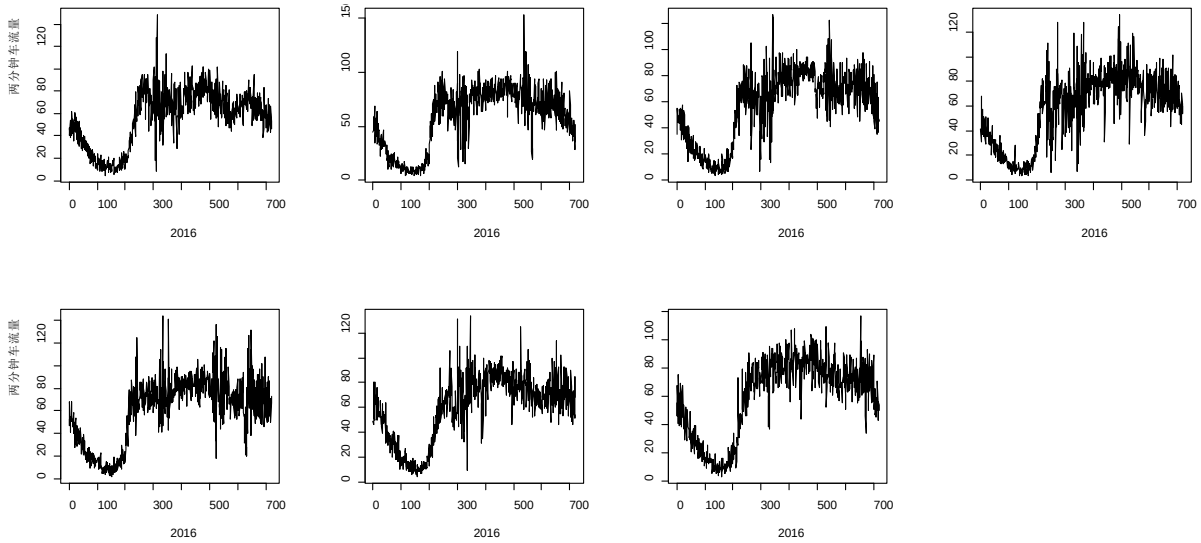

Figure 2. Guiyang Baoshan North Road and East New Road intersection south to north one week traffic flow
图 2. 贵阳市宝山北路与东新区路交叉口南向北一周车流量时序图

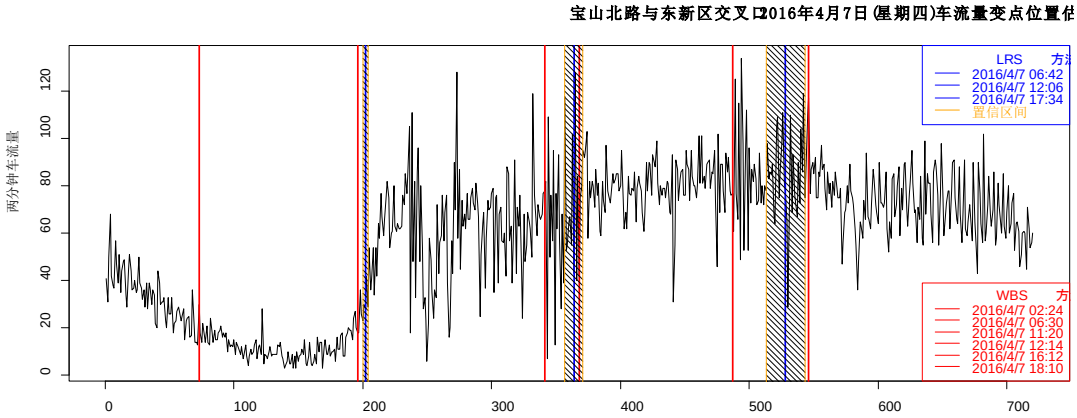

Figure 3. Intersection of Baoshan North Road and East New Roads Thursday, April 7, 2016 estimated traffic flow change point location
图 3. 宝山北路与东新区路交叉口 2016 年 4 月 7 日(星期四)车流量变点位置估计

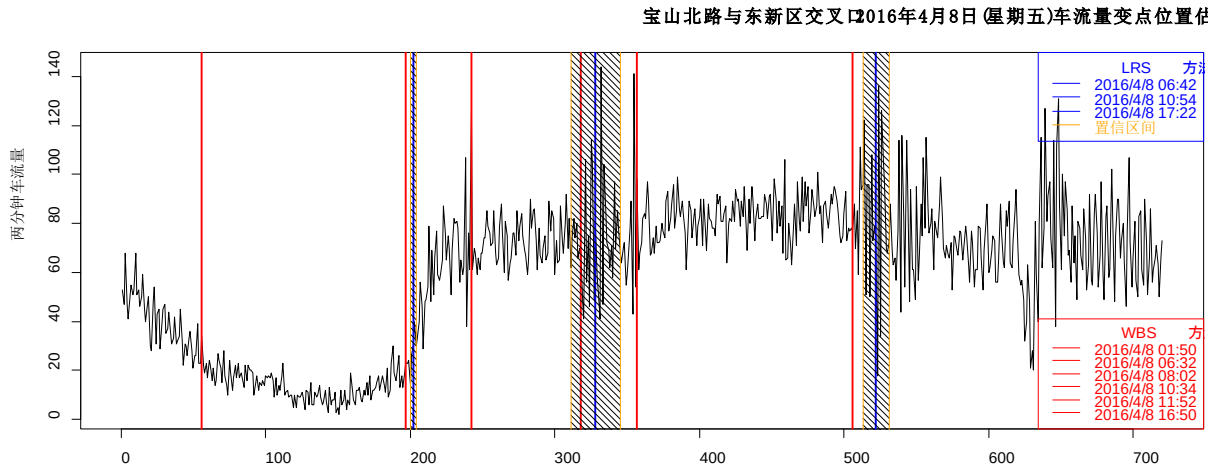

Figure 4. Intersection of Baoshan North Road and East New Roads Friday, April 8, 2016 estimated traffic flow change point location

图 4. 宝山北路与东新区路交叉口 2016 年 4 月 8 日(星期五)车流量变点位置估计

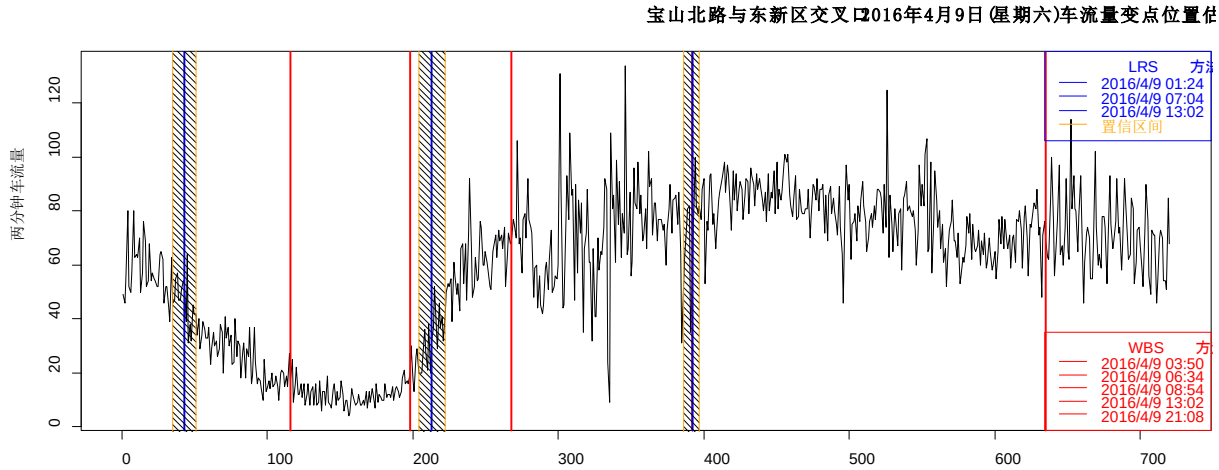

Figure 5. Intersection of Baoshan North Road and East New Roads Saturday, April 9, 2016 estimated traffic flow change point location

图 5. 宝山北路与东新区路交叉口 2016 年 4 月 9 日(星期六)车流量变点位置估计

通过图 3~5 可直观清晰的对比两种方法对变点位置的估计情况，WBS 方法检测出的变点数目明显多于 LRS 方法，过估计问题依然普遍存在。从工作日(星期四星期五)的变点位置不难虽然 WBS 方法估计变点多于 LRS 方法，但是 LRS 方法检测出的变点附近 WBS 方法也识别到了变点的存在，即对于变点的存在性检测两种方法都是足够灵敏的。区别较大的是休息日(星期六)的变点位置估计，LRS 方法和 WBS 方法检测出的点分别对应时间 01:24、07:04、13:02 和 03:50、06:34、08:54、13:02、21:08。但从实际出发，该交叉路口处于贵阳市中心，很多大型商圈围绕，当星期五结束工作，晚上人们会集中前往娱乐消遣，与朋友聚会等，所以在 01:24 左右前还属于人群活动时间，成为新的出行“高峰”，01:24 左右后人们才陆续休息，此时车流极具下降造成分布改变；休息日不存在早晚高峰问题，但考虑早晨人们开始起床活动或出游等造成 07:04 左右时交通流分布发生改变，车流量开始增多；非工作日贵阳车辆不实行尾号限行，所以全天车流都较多，在 13:02 后该路口车流量趋于平稳，分布不在发生改变。

表 4 展示了 LRS 方法和 WBS 方法对 2016 年 4 月 4 日(星期一)至 2016 年 4 月 10 日(星期日)每天车流量的变点识别情况，具体见表 4。

Table 4. Contrast of change point of traffic flow data detection results
表 4. 交通流数据变点检测结果对比

日期 方法	2016/4/4 (星期一)		2016/4/5 (星期二)		2016/4/6 (星期三)		2016/4/7 (星期四)		2016/4/8 (星期五)		2016/4/9 (星期六)		2016/4/10 (星期日)	
	位置	时间	位置	时间	位置	时间	位置	时间	位置	时间	位置	时间	位置	时间
LRS	226	07:30	201	06:40	220	07:18	202	06:42	202	06:42	43	01:24		
	345	11:28	328	10:54	339	11:16	364	12:06	328	10:54	213	07:04	无	
	568	18:54	537	17:52	534	17:46	528	17:34	522	17:22	392	13:02		
WBS	44	01:26	56	01:50	45	01:28	73	02:24	56	01:50	116	03:50	48	01:34
	200	06:40	196	06:30	211	07:00	196	06:30	197	06:32	198	06:34	209	06:56
	402	13:22	297	09:52	342	11:22	341	11:20	242	08:02	268	08:54	650	21:38
			328	10:54	432	14:22	368	12:14	318	10:34	392	13:02		
			532	17:42	527	17:32	487	16:12	357	11:52	635	21:08		
			566	18:50			546	18:10	506	16:50				

结果显示工作日(星期一至星期日) LRS 方法普遍识别出 3 个变点，这 3 个点将车流量数据分为 4 个子序列，变点位置及对应的时间也相对较为固定，集中在早、中、晚上下班高峰期，与实际也比较符合。针对星期日车流数据 LRS 方法未识别到，原因可能为：1) 当天数据比较特殊，2) 星期日贵阳城区不限号，当日在所有时间段内车流分布都基本保持不变，尽管从凌晨到上午 7 点这段时间过车量较少，但并未导致整个车流分布发生改变。其次，WBS 方法对变点的识别也是灵敏的，但是过估计问题也十分严重，常将一些奇异点也当作变点处理。通过数值模拟和实证分析都表明相较于目前使用较多的 WBS 方法，LRS 方法不仅能有效识别到变点，而且能够更为准确的对变点位置进行估计，说明 LRS 方法对变点研究是实践可用的。

6. 结束语

文章考虑了分段平稳时间序列的变点检测问题，假设观测值服从参数不同的非线性时间序列模型(AR 模型)，且变点数量及位置均未知，通过构建 LRS 统计量对序列进行检测，实现序列变点的初步识别，进一步结合 MDL 准则进行模型选择，估计变点的数量、位置，最后对每个估计变点分别构造置信区间。同时利用 WBS 算法对序列变点数目及位置进行估计，并对两种算法进行模拟比较。数值模拟结果显示，两种方法均对变点有良好的识别效果，但 LRS 方法明显优于 WBS 方法，LRS 方法的模型选择过程能更有效的剔除异常值，使得变点估计更为准确。最后，结合贵阳市中心道路车流量数据实例分析，表明方法对于交通流突变分析效果较好。

基金项目

贵州大学 2017 年研究生创新基金项目(研理工 2017067)；国家自然科学基金项目(11661018，11361015)；全国统计科学研究项目(2014LZ46)；贵州省自然科学基金项目(黔科合 J 字[2014]2058 号)。

参考文献

[1] Hinkley, D.V. (1970) Inference about the Change-Point in a Sequence of Random Variables. *Biometrika*, **57**, 1-17. https://doi.org/10.1093/biomet/57.1.1

[2] Hinkley, D.V. and Hinkley, E.A. (1970) Inference about the Change-Point in a Sequence of Binomial Variables. *Bio-*

metrika, **57**, 477-488. https://doi.org/10.1093/biomet/57.3.477

[3] Yao, Y.C. (1987) Approximating the Distribution of the Maximum Likelihood Estimate of the Change-Point in a Sequence of Independent Random Variables. *Annals of Statistics*, **15**, 1321-1328. https://doi.org/10.1214/aos/1176350509

[4] Picard, D. (1985) Testing and Estimating Change-Points in Time Series. *Advances in Applied Probability*, **17**, 841-867. https://doi.org/10.2307/1427090

[5] Davis, R.A. and Rodriguez-Yam, G.A. (2006) Structural Break Estimation for Nonstationary Time Series Models. *Journal of the American Statistical Association*, **101**, 223-239. https://doi.org/10.1198/016214505000000745

[6] Fryzlewicz, P. (2014) Wild Binary Segmentation for Multiple Change-Point Detection. *The Annals of Statistics*, **42**, 2243-2281. https://doi.org/10.1214/14-AOS1245

[7] Yau, C.Y. and Zhao, Z. (2015) Inference for Multiple Change Points in Time Series via Likelihood Ratio Scan Statistics. *Journal of the Royal Statistical Society*, **2015**, 78. http://doi.org/10.1111/rssb.12139

[8] Ling, S. (2016) Estimation of Change-Points in Linear and Nonlinear Time Series Models. *Econometric Theory*, **32**, 402-430. https://doi.org/10.1017/S0266466614000863

[9] Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2008) Break Detection for a Class of Nonlinear Time Series Models. *Journal of Time Series Analysis*, **29**, 834-867. https://doi.org/10.1111/j.1467-9892.2008.00585.x

[10] Yao, Y.C. (1984) Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches. *Annals of Statistics*, **12**, 1434-1447. https://doi.org/10.1214/aos/1176346802

[11] Jackson, B., Scargle, J.D., Barnes, D., *et al.* (2005) An Algorithm for Optimal Partitioning of Data on an Interval. *IEEE Signal Processing Letters*, **12**, 105-108. https://doi.org/10.1109/LSP.2001.838216