Hans Journal of Data Mining 数据挖掘, 2024, 14(2), 102-115
Published Online April 2024 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2024.142009

Hans汉斯

INAR模型的多变点探测

陈加琪，卢飞龙*

辽宁科技大学理学院，辽宁 鞍山

收稿日期：2024年3月11日；录用日期：2024年4月11日；发布日期：2024年4月22日

摘　要

本文考虑了分段平稳的整数值自回归模型的多变点问题。借助最小描述长度函数得到似然比扫描方法，并将似然比扫描方法运用到分段平稳的整数值自回归模型中。此外，还研究了当模型系数之和趋于1时的情况，此时针对似然比扫描方法中的最小描述长度函数进行了调整，以提高变点探测的准确性。然后通过大量的数值模拟，验证了似然比扫描方法在不同的模型参数设置下的有效性，最后并将其运用于一组精神分裂症患者在知觉速度测试中的日常观察得分数据的实证分析之中。

关键词

INAR模型，变点探测，似然比扫描方法

Multiple Change Point Detection for INAR Model

Jiaqi Chen, Feilong Lu*

College of Science, University of Science and Technology Liaoning, Anshan Liaoning

Received: Mar. 11th, 2024; accepted: Apr. 11th, 2024; published: Apr. 22nd, 2024

Abstract

This paper considers the multiple change-points problem in piecewise stationary integer-valued autoregressive model. Using the Minimum Description Length function, the likelihood ratio scanning method is obtained and applied to the piecewise stationary integer-value autoregressive model. In addition, when the sum of the model coefficients tends to 1, the Minimum Description Length function in the likelihood ratio scanning method is adjusted to improve the accuracy of the change point detection. Then, through many numerical simulations, the effectiveness of the like-

*通讯作者。

文章引用: 陈加琪, 卢飞龙. INAR 模型的多变点探测[J]. 数据挖掘, 2024, 14(2): 102-115.
DOI: 10.12677/hjdm.2024.142009

lihood ratio scanning method in different model parameter settings was verified. Finally, it was applied to the empirical analysis of the daily observations of the score achieved by a schizophrenic patient on a test of perceptual speed.

Keywords

INAR Model, Change Point Detection, Likelihood Ratio Scanning Method

1. 引言

在时间序列分析过程中，所关注的某些量可能在某个时刻发生显著变化，这个变化时刻被称作变点。如果在发现变点后仍旧基于之前的模型进行统计分析，就可能得到错误的分析结果，为了降低错误率，提升效益，因此有必要进行变点统计分析。最早的变点研究可以追溯到 20 世纪 50 年代，Page [1]针对质量控制问题发表了一篇关于变点探测的文章，自此变点问题在统计领域得到了广泛的讨论。目前，经过研究人员对变点方法不断地发展与完善，变点的处理方法已经有很多，部分方法也很成熟了。多变点探测常用的方法是通过优化特定的目标函数来寻找变点，如最小二乘法[2]、Bayes 方法[3]等。除此之外还有非参数方法，如 U-统计量法[4]、Wilcoxon 秩和法[5]、Kolmogorow 检验法[6]、累积和(Cumulative Sum, CUSUM)检验法[7] [8]。需要注意的是，变点组合的数量会随着样本量的增长而呈指数增长，这使得在优化过程中的计算难度变大。而为了避免优化难题，二分法是一个比较受欢迎的解决办法。二分法最开始是由 Vostrikova [9]和 Bai [10]提出并研究的，二分法更多地应用于均值多变点[11]和方差多变点[12]的探测中。Fryzlewicz [13]将二分法扩展为 Wild 二分法，该拓展方法非常重要，被命名为野生二进制分割(Wild Binary Segmentation, WBS)。除此之外，还有 Davis 等人[14]提出的遗传算法，Killick 等人[15]提出的截断精确线性时间(Pruned Exact Linear Time, PELT)方法以及 Yau 和 Zhao [16]提出的似然比扫描方法(Likelihood Ratio Scan Method, LRSM)。与 WBS，PELT 等方法相比，LRSM 不仅降低了计算复杂度，而且提高了变点估计的精度。

近年来兴起了对整数值时间序列的变点问题的研究。随着此类时间序列的几个相应模型的提出，整数值时间序列在各个领域中有着越发广泛的应用，例如在金融(在一个时间周期内事件的发生的概率)、气候学、医学等。关于此类时间序列相关模型的构建，Mckenzie [17]和 Al-osh 和 Alzaid [18]提出了一个带有泊松边际分布的 INAR(1)过程，即众所周知的 PINAR(1)过程，由于其简单的形式非常受欢迎。Alzaid 和 Alzaid [19]以及 Du 和 Li [20]将 INAR(1)过程进一步扩展成更一般的 INAR(p)过程。尽管 INAR 模型在实践中得到的应用非常广泛，但在拟合非线性现象时往往会出现不足。例如针对疫情出现拐点的这样的数据，平稳的 INAR 模型可能无法提供良好的拟合，需要针对变点对其进行分段拟合。因此，对于此类整数值时间序列，研究数据中的变点位置和数量，具有重要的实际意义。目前，据我们所知，关于此类整数值时间序列的变点探测的研究尚不多见，更多的是基于此类序列的变点检验的研究，例如文献[21] [22] [23]。

在现有的变点探测方法中，LRSM 简单且易于实施，并且当数据量为 n 时，计算复杂度从 n^2 降低到了 $\log(n)$，Yau 和 Zhao [16]在模拟研究中将 LRSM 与其他方法进行了对比，总体而言 LRSM 的适用性

更强并且比其他方法更准确些。因此，本文选择 LRSM 探测由分段平稳的 INAR 模型所产生的整数值时间序列数据中的多变点。首先基于分段平稳 INAR(1)过程给出变点问题的基本假设和似然比扫描方法的具体实施步骤。其次通过大量的数值模拟并结合实证分析证明 LRSM 的有效性。

本文的其余部分组织如下。第 2 节详细介绍了 INAR 模型以及 LRSM 的三个步骤。第 3 节通过大量的数值模拟，说明将似然比扫描方法应用于整数值自回归过程中的有效性。第 4 节将 LRSM 方法运用到一个真实案例中。最后，在第 5 节对全篇内容做出总结，概述本论文所做的工作。

2. INAR 模型以及似然比扫描方法

2.1. INAR 模型

在许多领域中，例如，经济学、医学、精算统计等，诸多有趣的变量都是非负整数值的，具体的有每次绿灯时路口通过的车辆数量、某城市每日发生抢劫案件的次数、每月某航班的起降次数、某个病人在服药前后的每日发病次数。因此近年来，关于非负整数值时间序列的建模受到越来越多的关注。此类数据相关模型的构建，最常用的是手段就是利用稀疏算子进行建模。Steutal 和 Van Harn [24]最早提出了二项稀疏算子“∘”，其定义如下：

$$\alpha \circ Y \triangleq \sum_{j=1}^{Y} z_j,$$

其中 $\alpha \in [0,1)$，Y 是非负的整数值随机变量，$\{z_j\}$ 为独立同分布的以 α 为参数的伯努利随机变量序列，并独立于 Y，其分布律为 $P\{z_j=1\}=1-P\{z_j=0\}=\alpha$。Du 和 Li [20]利用算子提出了 INAR(p)模型，这也是本文考虑的模型，其具体定义如下：

$$Y_t = \alpha_1 \circ Y_{t-1} + \cdots + \alpha_p \circ Y_{t-p} + \varepsilon_t,\ t \geq p,$$

其中 ε_t 是独立同分布(*i.i.d*)的非负整数值随机变量序列，且服从均值为 λ 的泊松分布。

关于该模型的的基本性质有：

1) Y_t 的条件均值：$E\left[Y_t \middle| Y_{t-1},\cdots,Y_{t-p}\right] = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \lambda$；

2) Y_t 的条件方差：$Var\left[Y_t \middle| Y_{t-1},\cdots,Y_{t-p}\right] = \sum_{j=1}^{p} \alpha_j \left(1-\alpha_j\right) Y_{t-j} + \lambda$；

3) Y_t 的均值：$E\left(Y_t\right) = \lambda \Big/ 1 - \sum_{j=1}^{p} \alpha_j$；

4) 当 $\alpha_j \in [0,1)$，$j=1,2,\cdots,p$，$\sum_{j=1}^{p} \alpha_j < 1$时，则模型是平稳遍历的。

2.2. 似然比扫描方法

LRSM 相比较其它的变点方法，可以快速有效的探测出时间序列中的变点。此方法主要有三个步骤：1) 扫描由分段平稳的 INAR 过程所产生的观测序列以获得初始的变点估计；2) 利用 MDL 函数进行 INAR 模型的选择过程，得到一致的变点估计；3) 对第二步中得到的变点集合 $\hat{J}^{(2)}$ 中的变点进行最终估计，以获得变点近似阶数达到最优时的变点估计。下面对基于 LRSM 探测分段平稳 INAR 过程中变点的三个步骤进行详细介绍。

为更好研究分段平稳 INAR 过程中的多变点问题，下面先给出一些基本设定。假如有观测序列 $\{X_t\}_{t=1,\cdots,n}$ 可以被分割成 $m+1$ 段平稳的 INAR 过程。设 $j=1,\cdots,m$，j 表示变点个数，用 τ_j 表示序列从第 j

段突然变化到第 $j+1$段变点的位置(第 j 个变点的位置)。令 $\tau_0 \triangleq 0$，$\tau_{m+1} \triangleq n$，则有 INAR 序列的第 j 段可以表示为：

$$Y_{t,j} = X_t, \quad \tau_{j-1} < t \le \tau_j, \quad j = 1,2,\cdots,m+1$$

其中 $\{Y_{t,j}\}$ 是一个平稳的 INAR 过程，满足：

$$Y_{t,j} = \alpha_{j,1} \circ Y_{t-1,j} + \cdots + \alpha_{j,p_j} \circ Y_{t-p_j,j} + \varepsilon_{t,j},$$

其中算子“∘”为二项稀疏算子，$\varepsilon_{t,j}$ 为独立同分布的非负整数值随机序列。第 j 段对应的 INAR(p_j)模型的参数向量为 $\boldsymbol{\theta}_j = (\alpha_{j,1},\cdots,\alpha_{j,p_j},\lambda_j)$。

假设 1 假设 $\sum_{i=1}^{p_j}\alpha_{j,i} < 1$，以保证每一段 INAR 过程 $\{Y_{t,j}\}_{j=1,2,\cdots,m+1}$ 都是平稳的。

假设 2 假设观测序列 $\{X_t\}_{t=1,\cdots,n}$ 被分割每一段 INAR 过程 $\{Y_{t,j}\}_{j=1,2,\cdots,m+1}$ 的阶数都是有限的，这里把每一段 INAR 模型的最大阶数用 $p_{\max}$ 来表示。

假设 3 这里用 $J = (\tau_1,\cdots,\tau_m)$，表示一系列的变点。定义各个变点的相对位置 $\{\varpi_j\}$，满足 $0 = \varpi_0 < \varpi_1 < \cdots < \varpi_m < \varpi_{m+1} = 1$ 且 $\tau_j = \lfloor \varpi_j n \rfloor$，这里 $\lfloor \varpi_j n \rfloor$ 表示 $\varpi_j n$ 的整数部分。假设存在 $\varepsilon_\varpi = \varepsilon_\varpi(n) > 0$，使得 $\min_{j=0,\cdots,m-1}(\varpi_{j+1} - \varpi_j) > \varepsilon_\varpi$。同时为了保证模型的可识别性，进一步假设 $n\varepsilon_\varpi > p_{\max}$。

接下来详细介绍运用 LRSM 探测分段平稳 INAR 过程中多变点的三个步骤。

第一步：扫描来自于分段平稳 INAR 过程的观测值 $\{X_t\}_{t=1,\cdots,n}$ 序列，得到最初估计的变点。

定义 t 时刻的扫描窗口为：

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

以及对应的观测值为：

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

其中 $t = h,\cdots,n-h$，h 为扫描窗口半径。扫描窗口半径 h 的选择与样本量 n 以及变点之间的距离有关，关于 h 的选择，详细可见第 3 节。

为了获得扫描窗口中最初估计的变点，可以选择似然比统计量。对于一个来自于 INAR(p)模型的样本 $D = (d_1,\cdots,d_n)$，则有 INAR 模型的条件似然函数如下：

$$L(\boldsymbol{\theta}) = \log \sum_{t=p+1}^{n} \left[P\left(d_t \middle| d_{t-1},\cdots,d_{t-p}\right) \right] \tag{1}$$

其中，

$$P\left(d_t \middle| d_{t-1},\cdots,d_{t-p}\right) = \sum_{i_1=0}^{\min(d_{t-1},d_t)} \binom{d_{t-1}}{i_1} \alpha_1^{i_1} (1-\alpha_1)^{d_{t-1}-i_1} \sum_{i_2=0}^{\min(d_{t-2},d_t-i_1)} \binom{d_{t-2}}{i_2} \alpha_2^{i_2} (1-\alpha_2)^{d_{t-2}-i_2} \cdots$$
$$\sum_{i_p=0}^{\min\left[d_{t-p},d_t-(i_1+\cdots+i_{p-1})\right]} \binom{d_{t-p}}{i_p} \alpha_p^{i_p} (1-\alpha_p)^{d_{t-p}-i_p} \frac{\mathrm{e}^{-\lambda}\lambda^{d_t-(i_1+\cdots+i_p)}}{\left[d_t-(i_1+\cdots+i_p)\right]!},$$

是在 d_t 之前给定观测值的转移概率。然后扫描窗口 $W_t(h)$ 的扫描统计量可以定义为：

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

其中 $L_{1h}\left(t,\hat{\theta}_1\right)$，$L_{2h}\left(t,\hat{\theta}_2\right)$ 和 $L_{\cdot h}\left(t,\hat{\theta}\right)$ 的定义与等式(1)中的 $L(\boldsymbol{\theta})$ 类似，但是使用的观测值是 $\{X_s\}_{t-h+1,\cdots,t}$，$\{X_s\}_{t+1,\cdots,t+h}$ 以及 $\{X_s\}_{W_t(h)}$，$\hat{\theta}_1$，$\hat{\theta}_2$ 以及 $\hat{\theta}$ 是基于序列 $\{X_s\}_{t-h+1,\cdots,t}$，$\{X_s\}_{t+1,\cdots,t+h}$ 以及 $\{X_s\}_{W_t(h)}$ 的参数 $\boldsymbol{\theta}$ 的估

计量。接下来，利用 $S_h(t)$ 扫描观测到的序列，可以得到一系列的似然比扫描统计量值的序列 $\left(S_h(h), S_h(h+1), \cdots, S_h(n-h)\right)$，假如 t 是变点，则 $S_h(t)$ 会趋于很大。如果选择的 h 满足 $2h < n\varepsilon_\varpi$ 以及 $h > p_{\max}$，则每个扫面窗口中最多存在一个真实变点。因此，通过对局部变点位置的估计，可以得到潜在变点集合并定义为：

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

其中当 $t < h$ 且 $t > n-h$ 时，$S_h(t) \triangleq 0$。如果以点 m 为中心的窗口 $[m-h+1, m+h]$ 中的最大值为 $S_h(m)$，则 m 为一个局部变点。则有一系列的初始变点 $\hat{J}^{(1)} = \left(\hat{\tau}_1^{(1)}, \hat{\tau}_2^{(1)}, \cdots, \hat{\tau}_{\hat{m}^{(1)}}^{(1)}\right)$，其中 $\hat{m}^{(1)} = \left|\hat{J}^{(1)}\right|$。

定理 1 记真实变点集合为 $J_0 = \left(\tau_1^0, \tau_2^0, \cdots, \tau_{m_0}^0\right)$，假设 $2h < n\varepsilon_\varpi$ 且 $\varepsilon_\varpi > cn^{-q}$，则存在 $q \in [0,1)$，$c > 0$ 满足：

$$P\left(\max_{\tau_i^0 \in J_0} \min_{\hat{\tau}_j^{(1)} \in \hat{J}^{(1)}} \left|\tau_i^0 - \hat{\tau}_j^{(1)}\right| < h \right) \to 1.$$

定理 1 表明全部真实变点在初始估计变点集合 $\hat{J}^{(1)}$ 的 h-领域内。并且当变点之间的最小距离为 $n\varepsilon_\varpi = O\left(n^{1-q}\right)$ 时，估计的变点个数 $\hat{m}^{(1)}$ 以 $O\left(n^q\right)$ 的速度随着样本量的增加而增加。但此时并不能保证 $\hat{m}^{(1)}$ 等于真实的变点个数 m_0。也就是说，在第一步中变点的数量可能会被高估。因此接下来通过第二步，即利用合适的信息准则从 $\hat{J}^{(1)}$ 中再次筛选，以获得一致的变点估计。

第二步：优化目标函数，选择 INAR 模型，得到一致的变点估计。

为了从 $\hat{J}^{(1)}$ 中找出准确的变点，可以使用最小描述长度准则[25] (简记 MDL)进行 INAR 模型的选择过程。定义 MDL 如下：

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

其中，$J = (\tau_1, \cdots, \tau_m)$ 的取值是第一步中得到的初始估计变点集合 $\hat{J}^{(1)}$，$(n_1, \cdots, n_{m+1})$ 和 $P = (p_1, \cdots, p_{m+1})$ 分别是所有片段的长度和相应的 INAR 片段的阶数，$L_j\left(\hat{\theta}_j\right)$ 是第 j 个 INAR 片段的似然统计量。通过 MDL 准则的优化后，可以得到优化后的变点集合为：

$$\left(\hat{m}^{(2)}, \hat{J}^{(2)}, \hat{P}^{(2)}\right) = \underset{\substack{m=|J|, J \subseteq \hat{J}^{(1)} \\ P \in \{1, \cdots, p_{\max}\}^m}}{\arg\min} MDL(m, J, P),$$

其中 $\hat{J}^{(2)} = \left(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)}, \cdots, \hat{\tau}_{\hat{m}^{(2)}}^{(2)}\right)$，而 $\hat{m}^{(2)} = \left|\hat{J}^{(2)}\right|$。在第二步中得到的变点估计具有如下性质：

定理 2 在定理 1 的成立的条件下，令 $q = 0$，则有 $\hat{m}^{(2)} \xrightarrow{p} m_0$。并且，当 $\hat{m}^{(2)} = m_0$，则有：

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

$$\max_{j=1,\cdots,m_0+1} \left|\hat{p}_j^{(2)} - p_j^0\right| \xrightarrow{p} 0.$$

根据定理 2 可知变点 $\tau_j^0 \in \left[\hat{\tau}_j^{(2)} - h + 1, \hat{\tau}_j^{(2)} + h\right]$。由于 $h \geq d\log(n) \to \infty$，因此有 $\max_{j=1,\cdots,\hat{m}^{(2)}} \left|\hat{\tau}_j^{(2)}\right| = O_p(h)$ 与 $O_p(1)$ 相比的话得到的变点近似阶数并不是最优的，这里 d 的取值可参考文献[16]。通过运用文献[26]中的思想来获得变点位置的最终估计使其近似的阶数达到最优。

第三步：获得使得变点近似阶数达到最优的最终的变点估计。

定义第 j 个估计变点 $\hat{\tau}_j^{(2)}$ 的拓展的局部窗口和相应的观测值分别为：

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

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

令

$$L_j(\tau,\theta_1,\theta_2)=\sum_{t=\hat{\tau}_j^{(2)}-2h+1}^{\tau}l_t(\theta_1)+\sum_{t=\tau+1}^{\hat{\tau}_j^{(2)}+2h}l_t(\theta_2),$$

对于 $j=1,\cdots,\hat{m}^{(2)}$，定义变点最后的估计量为：

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

其中 $\hat{\theta}_j=\hat{\theta}_j(\tau)=\arg\max_{\theta_1}\sum_{t=\hat{\tau}_j^{(2)}-2h+1}^{\tau}l_t(\theta_1)$，同理可以得到 $\hat{\theta}_{j+1}$ 的定义。同样地，第 j 段的 INAR 模型的阶数取 $\hat{p}_j^{(2)}$。

定理 3 在定理 2 成立的条件下，若 $3h<n\varepsilon_\varpi$，则有：

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

且

$$\hat{\tau}_j^{(3)}-\tau_j^0\xrightarrow{d}\underset{\tau_j}{\arg\max}W\left(\tau_j,\theta_j^0,\theta_{j+1}^0\right),$$

其中 $W(\tau_j):=W\left(\tau_j,\theta_j^0,\theta_{j+1}^0\right)$ 是一个双边随机游走如下所示：

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

定理 1~3 证明过程可参考文献[16]。

3. 模拟研究

在本节中，通过大量的数值模拟来评估 LRSM 在由分段平稳的 INAR 过程所产生的有限样本下变点识别性能，所有模拟实验均通过 R 语言程序实现。当样本量小于 800 时，可以考虑 $h=\max\left\{25,(\log n)^2\right\}$，当样本量 n 大于 800 时，可以考虑 $h=\max\left\{50,2(\log n)^2\right\}$。在本文中，样本量大于 800 时，取 $h=2(\log n)^2$；如果样本量小于 800 时，取 25。所有模拟结果都经过 100 次循环得到。关于变点数量以及位置 $\tau_j=\left[\varpi_j n\right]$ 的设定，分别设置成没有变点，有一个变点，有两个变点三种情况。在只有一个变点的情况下 ϖ_1 取 0.5；在有两个变点的情况下，$\{\varpi_1,\varpi_2\}=\{0.34,0.68\}$ 或者 $\{\varpi_1,\varpi_2\}=\{0.39,0.68\}$。通过计算频率 $f_0=\sum_{i=1}^{100}I_{\hat{m}_i^{(2)}=m_0}$ 来检验 LRSM 是否可以准确探测出变点数量，其中 m_0 表示真正的变点数量。同时为了评估 LRSM 估计精度，给出如下定义：如果一个方法若能准确的探测出变点数量，并且当 $n=1024$ 时，每对真实变点和估计变点之间的距离在 50 以内；当 $n=120$ 时，每对真实变点和估计变点之间的距离在 5 以内，认为探测出来的变点是有效的。在 $\hat{m}_i^{(2)}=m_0$ 时，还通过计算均值，和均方误差来评估估计精度。

LRSM 的三个步骤中都需要求对相应的模型参数的极大似然估计值，但在实际操作中采用的是矩估计值的绝对值，具体原因如下：1) 众所周知，极大似然估计法需要通过数值优化来求解，因此求解过程耗费时间长；矩估计因其有显著的表达式，求解过程简单，耗费时间短。2) 通过模拟研究发现，采用矩估计值几乎不影响 LRSM 的性能。3) 在模拟研究中，由于生成的随机数，具有一定的随机性，因此有时参数的矩估计值需要调整，而本文采用的调整手段是直接取矩估计值的绝对值。

3.1. INAR(1)的模拟研究

在本小节中，通过数值模拟来评估 LRSM 在由分段平稳的 INAR(1)过程所产生的有限样本下变点识别性能，所考虑的模型有：模型 A-D。模型 A-D 的样本路径图见图 1，模拟结果见表 1。

Table 1. Simulation results from Model A to Model D
表 1. 从模型 A 到模型 D 的模拟结果

模型		f_0 [1]	*exact*[2]	$\hat{\tau}$ [3]	*mean*[4]	*MSE*[5]
模型 A	$\alpha=0.1,\lambda=1$	100	-	-	-	-
	$\alpha=0.2,\lambda=2$	100	-	-	-	-
	$\alpha=0.3,\lambda=3$	98	-	-	-	-
	$\alpha=0.4,\lambda=4$	100	-	-	-	-
	$\alpha=0.5,\lambda=5$	100	-	-	-	-
	$\alpha=0.6,\lambda=4$	100	-	-	-	-
	$\alpha=0.7,\lambda=3$	100	-	-	-	-
	$\alpha=0.8,\lambda=2$	100	-	-	-	-
	$\alpha=0.9,\lambda=1$	98	-	-	-	-
模型 B		99	99	512	512.4	2.78
模型 C		98	98	400 612	402.11 612.16	15.79 28.76
模型 D		100	99	69	68.09	2.98

[1] 正确识别变点数量的次数；[2] 有效识别变点的次数；[3] 真实变点的位置；[4] 有效识别到的变点的平均值；[5] 有效识别到的变点的均方误差。

模型 A：没有变点的平稳 INAR(1)过程

模型 A 用来评估 LRSM 在没有变点时的性能，即当观测值中无变点时，该方法能否准确识别。设样本量 $n=1024$，并采用多种参数组合的 INAR(1)过程：$X_t=\alpha\circ X_{t-1}+\varepsilon_t$ 生成观测值，其中 ε_t 服从均值为 λ 的泊松分布，α 和 λ 的取值见表 1。从表 1 中可以看出，在 $\alpha=0.3$，$\lambda=3$ 以及 $\alpha=0.9$，$\lambda=1$ 时正确识别变点数量的次数只有 98 次，而其它参数组合的情况下，正确识别变点的次数都是 100 次。由于随机因数导致 LRSM 偶尔会有一两次错误的识别到变点。总的来看在没有变点的情况下，LRSM 是非常准确有效的。

模型 B：有一个变点的分段平稳 INAR(1)过程

$$X_t = \begin{cases} 0.5 \circ X_{t-1} + \varepsilon_{t,1}(1) & 1 \le t \le 512, \\ 0.5 \circ X_{t-1} + \varepsilon_{t,2}(4) & 513 \le t \le 1024. \end{cases}$$

其中 $\varepsilon_{t,j}(\lambda_j)$ 表示分段平稳的 INAR 模型的新息过程 $\varepsilon_{t,j}$ 服从均值为 λ_j 的泊松分布。下面模型 C-D 的 $\varepsilon_{t,j}(\lambda_j)$ 的含义与这里的 $\varepsilon_{t,j}$ 的一致。

模型 B 在 512 处有一个变点。从表 1 可以看出，LRSM 正确识别变点数量的频率高达 99 次，且全都满足了估计变点与真实变点之间的距离均小于 50；而估计变点的均值约为 512，且均方误差也是合理的。所以针对模型 B 这种情况，LRSM 的估计效果也很理想。

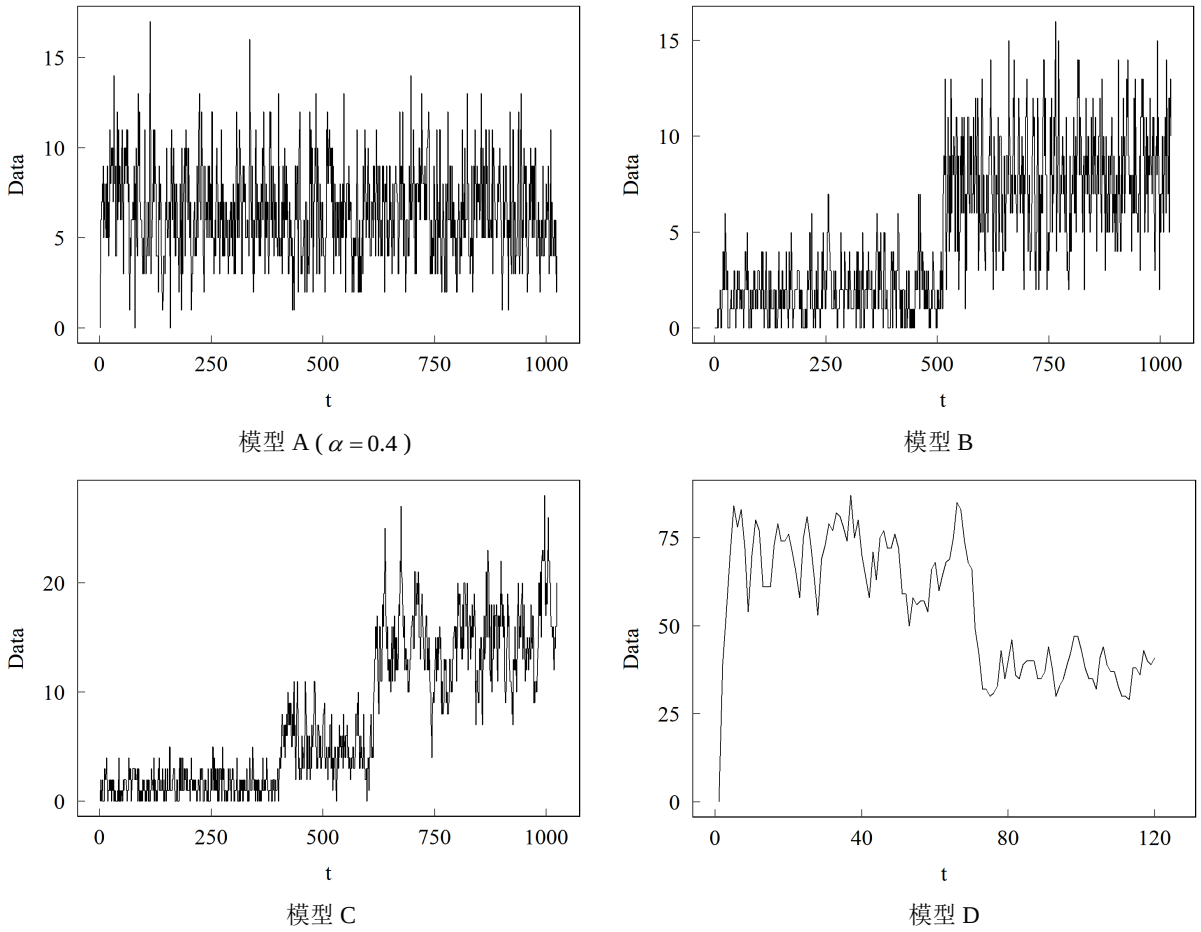

Figure 1. Sample path from Model A to Model D
图 1. 模型 A 到模型 D 的样本路径图

模型 C：有两个变点的分段平稳 INAR(1)过程

$$X_t = \begin{cases} 0.3 \circ X_{t-1} + \varepsilon_{t,1}(1) & 1 \le t \le 400, \\ 0.6 \circ X_{t-1} + \varepsilon_{t,2}(2) & 401 \le t \le 612, \\ 0.8 \circ X_{t-1} + \varepsilon_{t,3}(3) & 613 \le t \le 1024. \end{cases}$$

模型 C 中有两个变点，分别在 400 和 612。正确识别变点个数的次数有 98 次，而 98 次的结果都是有效的。探测出来的两个变点的均值皆接近真实值，均方误差也不是太大。但是随着变点个数的增加，

变点识别难度也会提升，LRSM 变点识别准确率有所降低，总的来看估计效果还是比较理想。

模型 D：在小样本情况下，有一个变点的分段平稳 INAR(1)过程

$$X_t = \begin{cases} 0.455 \circ X_{t-1} + \varepsilon_{t,1}(38.14) & 1 \le t \le 69, \\ 0.64 \circ X_{t-1} + \varepsilon_{t,2}(13.62) & 70 \le t \le 120. \end{cases}$$

模型 D 用于评估当样本量较小时，LRSM 的性能。所有参数以及变点位置均根据第 4 节中的真实案例分析设定。从表 1 中可以看出，无论是在准确探测变点数量上还是有效探测变点位置上，LRSM 都表现良好。

综上所述，LRSM 能较为高效准确的探测出 INAR(1)模型中的变点个数和位置。

3.2. INAR(2)的模拟研究

在本小节中，通过数值模拟来评估 LRSM 在由高阶的分段平稳 INAR 过程所产生的有限样本下变点识别性能，所考虑的模型有：模型 E~G。模型 E~G 的样本路径图见图 2，模拟结果见表 2。

Table 2. Simulation results from Model E to Model G
表 2. 从模型 E 到模型 G 的模拟结果

模型	f_0 [1]	*exact*[2]	$\tilde{\tau}$ [3]	*mean*[4]	*MSE*[5]
模型 E	100	-	-	-	-
模型 F	97	97	512	512.72	15.92
模型 G	97	97	350 700	351.35 699.41	9.1 5.7

[1] 正确识别变点数量的次数；[2] 有效识别变点的次数；[3] 真实变点的位置；[4] 有效识别到的变点的平均值；[5] 有效识别到的变点的均方误差。

模型 E：无变点的平稳 INAR(2)过程

$$X_t = 0.5 \circ X_{t-1} + 0.2 \circ X_{t-2} + \varepsilon_{t,1}(1) \quad 1 \le t \le 1024.$$

模型 F：具有一个变点的分段平稳 INAR(2)过程

$$X_t = \begin{cases} 0.5 \circ X_{t-1} + 0.1 \circ X_{t-2} + \varepsilon_{t,1}(1) & 1 \le t \le 512, \\ 0.3 \circ X_{t-1} + 0.1 \circ X_{t-2} + \varepsilon_{t,2}(4) & 513 \le t \le 1024. \end{cases}$$

模型 G：具有两个变点的分段平稳 INAR(2)过程

$$X_t = \begin{cases} 0.25 \circ X_{t-1} + 0.2 \circ X_{t-2} + \varepsilon_{t,1}(1) & 1 \le t \le 350, \\ 0.25 \circ X_{t-1} + 0.2 \circ X_{t-2} + \varepsilon_{t,2}(4) & 351 \le t \le 700, \\ 0.25 \circ X_{t-1} + 0.2 \circ X_{t-2} + \varepsilon_{t,3}(1) & 701 \le t \le 1024. \end{cases}$$

其中 $\varepsilon_{t,j}(\lambda_j)$ 表示分段平稳的 INAR 模型的新息过程 $\varepsilon_{t,j}$ 服从均值为 λ_j 的泊松分布。

由于在模型 INAR(2)中 LRSM 变点识别性能与在模型 INAR(1)中表现相似，因此这里只简单阐述上述三种模型的结果。从表 2 中可以看出，在模型 E 中，也就是在没有变点的情况下，LRSM 是完全准确和有效的。在模型 F 和模型 G 中，LRSM 有 97 次准确的识别出变点的数量，并且均满足估计出的变点与真实变点的距离小于 50。其中有效估计出的变点中，其变点位置的均值与真实变点之间均相差 1.5 以内。综上所述，LRSM 可以有效准确地探测出 INAR(2)模型变点的数量以及位置。

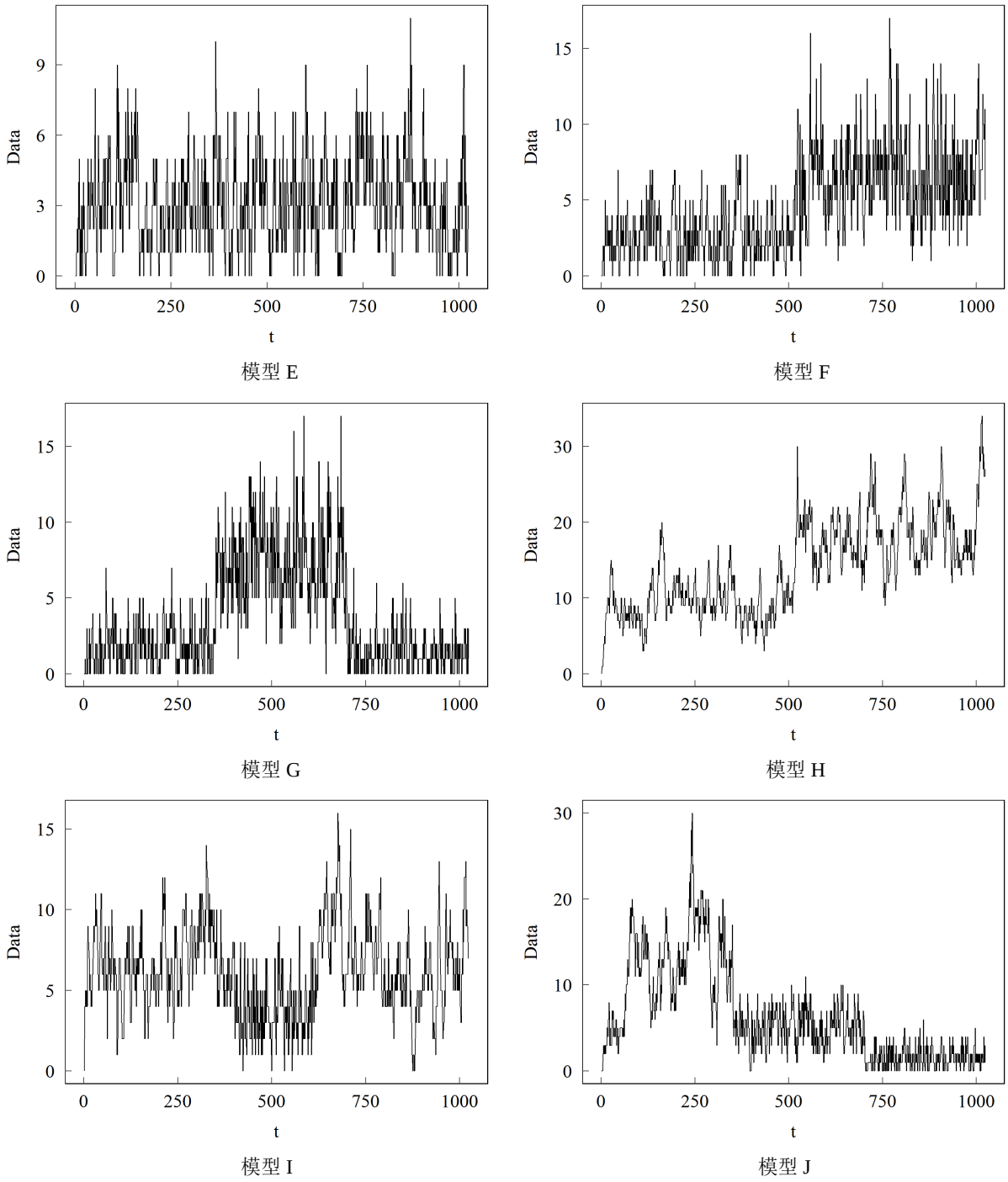

Figure 2. Sample path from Model E to Model J
图 2. 模型 E 到模型 J 的样本路径图

3.3. INAR 模型的系数之和接近于 1 时的模拟研究

当使用 LRSM 处理 INAR 模型中 $\sum_{i=1}^{p_j}\alpha_{j,i}\to 1$ 的变点问题时，发现 LRSM 在第二步偶尔会保留非变点的情况，使得最终估计的变点数量很容易大于真实的变点数量。第二步中使用的 MDL 准则来源于计算

理论，是 Rissanen [27]研究通用编码时提出的，并且证明了残差的编码长度由拟合模型的对数似然的负数给出。因此，本文对 MDL 准则进行了调整。也就是说，残差的编码长度由拟合模型的对数似然的负数给出，而不是条件对数似然的负数。调整后的 MDL 函数如下：

$$\mathrm{MDL}(m,J,P)=\log(m)+(m+1)\log(n)+\sum_{j=1}^{m+1}\log(p_j)+\sum_{j=1}^{m+1}\frac{p_j+1}{2}\log(n_j)-\sum_{j=1}^{m+1}L_j(\hat{\theta}_j)-\sum_{j=1}^{m+1}\hat{L}_j(\hat{\theta}_j),$$

其中 $\hat{L}(\boldsymbol{\theta})=\log P(d_1,d_2,\cdots,d_p)+L(\boldsymbol{\theta})$，$L(\boldsymbol{\theta})$的定义见式子。为了便于表述，调整后的 LRSM 被记录为全似然比扫描方法(简记 FLRSM)。在本小节中，通过把 FLRSM 用于由模型 H~J 生成的分段平稳整数值自回归多变点过程来评估 FLRSM 的性能，同时将其与以前的 LRSM 进行比较。模型 H~J 的样本路径图见图 2，模拟结果见表 3。

Table 3. Simulation results from Model H to Model J
表 3. 从模型 H 到模型 J 的模拟结果

模型	方法	f_0 [1]	*exact*[2]	$\hat{\tau}$ [3]		*mean*[4]		*MSE*[5]	
模型 H	LRSM	92	85	512		512.87		380.68	
	FLRSM	98	92	512		512.63		449.49	
模型 I	LRSM	95	94	400	612	405.52	607.49	200.65	116.38
	FLRSM	99	98	400	612	404.29	609.32	97.42	84.28
模型 J	LRSM	94	91	350	700	352.82	698.59	120.58	28.57
	FLRSM	96	96	350	700	352.06	698.59	75.56	21.2

[1] 正确识别变点数量的次数；[2] 有效识别变点的次数；[3] 真实变点的位置；[4] 有效识别到的变点的平均值；[5] 有效识别到的变点的均方误差。

模型 H：具有一个变点的分段平稳 INAR(1)过程

$$X_t=\begin{cases}0.9\circ X_{t-1}+\varepsilon_{t,1}(1) & 1\le t\le 512,\\ 0.9\circ X_{t-1}+\varepsilon_{t,2}(2) & 513\le t\le 1024.\end{cases}$$

模型 I：具有两个变点的分段平稳 INAR(1)过程

$$X_t=\begin{cases}0.85\circ X_{t-1}+\varepsilon_{t,1}(1) & 1\le t\le 400,\\ 0.5\circ X_{t-1}+\varepsilon_{t,2}(2) & 401\le t\le 612,\\ 0.85\circ X_{t-1}+\varepsilon_{t,3}(1) & 613\le t\le 1024.\end{cases}$$

模型 J：具有两个变点的分段平稳 INAR(2)过程

$$X_t=\begin{cases}0.83\circ X_{t-1}+0.1\circ X_{t-2}+\varepsilon_{t,1}(1) & 1\le t\le 350,\\ 0.49\circ X_{t-1}+0.1\circ X_{t-2}+\varepsilon_{t,2}(2) & 351\le t\le 700,\\ 0.25\circ X_{t-1}+0.1\circ X_{t-2}+\varepsilon_{t,3}(1) & 701\le t\le 1024.\end{cases}$$

从表 3 可以看出，数据波动越大，$\sum_{i=1}^{p_j}\alpha_{j,i}$ 越接近于 1，探测变点的难度就越大。在探测变点的数量和位置上，FLRSM 的性能优于 LRSM。但两种方法得到的均方误差都比较大，估计精度还有待提高。

4. 实证分析

在本节中将 LRSM 运用到真实的整数值时间序列数据中。在拟合具体模型时需要估计相应的参数值，主要考虑矩估计值和最大似然估计值。通过计算拟合模型预测序列的均方误差，本文最终选择了均方误差偏小的矩估计值。

这个例子以日常观察精神分裂症患者在感知速度测试中获得的分数为观测值，见 McCleary 和 Hay [28] 数据如图 3 所示。数据由 120 个连续的日常得分组成。然而从第 61 天起，患者开始注射一种强效镇静剂(氯丙嗪)，此药剂有望降低感知速度。Neal 和 Subba [29]猜测第 60 天是变点，因此通过两个不同的 INAR 模型拟合数据的两个部分。然而，他们没有提供任何正式估计变点的方法。Kashikar 等人[30]通过马尔可夫链蒙特卡罗(MCMC)程序，估计的变点位置在第 80 天。

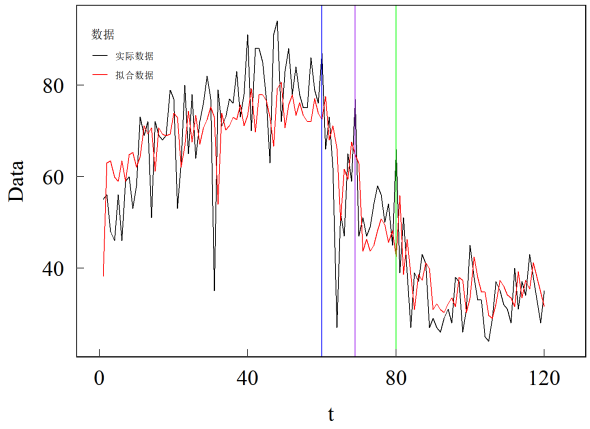

Figure 3. Schizophrenic actual data and fitted data
图 3. 精神分裂症的实际数据和拟合数据

针对该数据，首先运用 LRSM 探测出变点位置在第 69 天，并进行分段 INAR(1)模型拟合，最后估计了拟合模型的参数。同时注意到，在服用镇静剂后，患者评分开始下降，这是 Neal 和 Subba 猜测变点在第 60 天的基础。从评分图(图 3)可以看出，患者在 69 天后的评分明显低于 69 天前的患者评分。80 天后，患者评分下降更为明显。其中图中的黑色曲线表示实际数据，红色曲线表示模型拟合值。

这意味着镇静剂从使用的第一天开始起作用，到服药后第 9 天疗效更明显，20 天后，即第 80 天，药效才得到充分发挥。因此，把第 60 天、第 69 天、第 80 天看作变点是合理的。为此，本文对三个疑似变点分别进行分段的 INAR(1)模型拟合，并利用预测值序列 $\left\{\hat{X}_t\right\}$ 的均方误差来选择合适的变点，均方误差计算公式如下：

$$MSE=\frac{1}{120}\sum_{t=1}^{120}\left(X_t-\hat{X}_t\right)^2,$$

其中预测值序列 $\hat{X}_t=E\left(X_t\middle|X_{t-1}\right)$。通过计算得到，当变点为 60 天时，此时预测值的均方误差为 137.31；变点为第 69 天时相应的均方误差为 110.31；而变点为第 80 天时，模型预测值的均方误差为 111.81。这清楚地表明，当变点为第 69 天时，拟合的 INAR(1)模型的性能更好。同时也可以观察到图 3 所示的模型的拟合曲线支持变点为第 69 天时拟合的 INAR(1)模型。此时相应拟合模型为：

$$X_t=\begin{cases}0.455\circ X_{t-1}+\varepsilon_{t,1}(38.14) & 1\le t\le 69,\\ 0.64\circ X_{t-1}+\varepsilon_{t,2}(13.62) & 70\le t\le 120.\end{cases}$$

5. 小结

本文主要运用 LRSM 探测分段平稳的整数值时间序列中变点的数量和位置。此外，针对模型系数之和趋近于 1 的情况，提出了 FLRSM，有效提高了变点探测精度。最后，将 LRSM 应用于一个真实的生物特征数据集，即精神分裂症患者数据。该数据集里面存在一个变点，而 LRSM 能有效地探测出该变点位置。通过模拟研究和实际数据分析的结果表明：LRSM 能够有效地探测出由 INAR 模型生成的分段平稳整数值自回归过程中的多变点。

基金项目

辽宁科技大学博士启动资金(601010391)。

参考文献

[1] Page, E.S. (1955) A Test for a Change in a Parameter Occurring at an Unknown Point. *Biometrika*, **42**, 523-527. https://doi.org/10.1093/biomet/42.3-4.523

[2] Bauwens, L., Backer B. and Dufays, A. (2014) A Bayesian Method of Change-Point Estimation with Recurrent Regimes: Application to GARCH Models. *Journal of Empirical Finance*, **29**, 207-229. https://doi.org/10.1016/j.jempfin.2014.06.008

[3] Barry, D. and Hartigan, J.A. (1993) A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association*, **88**, 309-319. https://doi.org/10.1080/01621459.1993.10594323

[4] Csörgő, M. and Horváth L. (1988) 20 Nonparametric Methods for Changepoint Problems. *Handbook of Statistics*, **7**, 403-425. https://doi.org/10.1016/S0169-7161(88)07022-1

[5] Wilcox, R.R. (2005) Introduction to Robust Estimation and Hypothesis Testing. *Technometrics*, **40**, 77-78. https://doi.org/10.1080/00401706.1998.10485491

[6] Schmid, F. and Trede, M. (1998) A Kolmogorov-Type Test for Second-Order Stochastic Dominance. *Statistics & Probability Letters*, **37**, 183-193. https://doi.org/10.1016/S0167-7152(97)00116-8

[7] Kim, S., Cho, S. and Lee, S. (2000) On the Cusum Test for Parameter Changes in Garch (1,1) Models. *Communication in Statistics-Theory and Methods*, **29**, 445-462. https://doi.org/10.1080/03610920008832494

[8] Lee, S., Tokutsu, Y. and Maekawa, K. (2004) The Cusum Test for Parameter Change in Regression Models with ARCH Error. *Journal of the Japanese Statistical Society*, **34**, 173-188. https://doi.org/10.14490/jjss.34.173

[9] Vostrikova, L.Y. (1981) Detecting "Disorder" in Multidimensional Random Processes. *Soviet Mathematics Doklady*, **24**, 55-59.

[10] Bai, J. (1997) Estimating Multiple Breaks One at a Time. *Econometric Theory*, **13**, 315-352. https://doi.org/10.1017/S0266466600005831

[11] Inclan, C. and Tiao, G.C. (1994) Use of Cumulative Sums of Squares for Retrospective Detection of Change of Variance. *Journal of the American Statistical Association*, **89**, 913-923. https://doi.org/10.1080/01621459.1994.10476824

[12] Berkes, J., Gombay, E. and Horvath, L. (2009) Testing for Changes in the Covariance Structure of Linear Processes. *Journal of Statistical Planning and Inference*, **139**, 2044-2063. https://doi.org/10.1016/j.jspi.2008.09.004

[13] 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

[14] Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2005) Structural Breaks Estimation for Non-Stationary Time Series Signals. *EEE/SP* 13*th Workshop on Statistical Signal Processing*, Bordeaux, 17-20 July 2005, 233-238. https://doi.org/10.1109/SSP.2005.1628598

[15] Killick, R., Fearnhead, P. and Eckley, I.A. (2012) Optimal Detection of Change Points with a Linear Computational Cost. *Journal of the American Statistical Association*, **107**, 1590-1598. https://doi.org/10.1080/01621459.2012.737745

[16] Yau, C.Y. and Zhao, Z. (2016) Inference for Multiple Change Points in Time Series via Likelihood Ratio Scan Statistics. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **78**, 895-916. https://doi.org/10.1111/rssb.12139

[17] McKenzie, E. (1985) Some Simple Models for Discrete Variate Time Series[1]. *JAWRA Journal of the American Water Resources Association*, **21**, 645-650. https://doi.org/10.1111/j.1752-1688.1985.tb05379.x

[18] Al-Osh, M.A. and Alzaid, A.A. (1987) First Order Integer Valued Autoregressive (INAR(1)) Process. *Journal of Time*

Series Analysis, **8**, 261-275. https://doi.org/10.1111/j.1467-9892.1987.tb00438.x

[19] Alzaid, A.A. and Al-Osh, M.A. (1990) An Integer-Valued *p*th-Order Autoregressive Structure (INAR(*p*)) Process. *Journal of Applied Probability*, **27**, 314-324. https://doi.org/10.2307/3214650

[20] Jin-Guan, D. and Yuan, L. (1991) The Integer-Valued Autoregressive (INAR(*p*)) Model. *Journal of Time Series Analysis*, **12**,129-142. https://doi.org/10.1111/j.1467-9892.1991.tb00073.x

[21] Cui, Y. and Wu, R. (2020) Test of Parameter Changes in a Class of Observation-Driven Models for Count Time Series. *Communications in Statistics-Theory and Methods*, **49**, 1933-1959. https://doi.org/10.1080/03610926.2019.1565843

[22] Zhang, C., Wang, D., *et al.* (2022) Generalized Poisson Integer-Valued Autoregressive Processes with Structural Changes. *Journal of Applied Statistics*, **49**, 2717-2739. https://doi.org/10.1080/02664763.2021.1915255

[23] Yu, K., Wang, H. and Weiß, C.H. (2022) An Empirical-Likelihood-Based Structural-Change Test for INAR Processes. *Journal of Statistical Computation and Simulation*, **93**, 442-458. https://doi.org/10.1080/00949655.2022.2109635

[24] Steutel, F.W. and Harn, K. (1979) Discrete Analogues of Self-Decomposability and Stability. *The Annals of Probability*, **7**, 893-899. https://doi.org/10.1214/aop/1176994950

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

[26] 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

[27] Rissanen, J. (1988) Stochastic Complexity in Statistical Inquiry. World Scientific, Singapore.

[28] McCleary, R. and Hay, R.A., *et al.* (1980) Applied Time Series Analysis for the Social Sciences. Sage Publications, Beverly Hills.

[29] Neal, P. and Subba R.T. (2007) MCMC for Integer-Valued ARMA Processes. *Journal of Time Series Analysis*, **28**, 92-110. https://doi.org/10.1111/j.1467-9892.2006.00500.x

[30] Kashikar, A.S., Rohan, N. and Ramanathan, T.V. (2013) Integer Autoregressive Models with Structural Breaks. *Journal of Applied Statistics*, **40**, 2653-2669. https://doi.org/10.1080/02664763.2013.823920