Statistics and Application 统计学与应用, 2024, 13(2), 521-527
Published Online April 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.132052

Hans汉斯

基于CUSUM方法的百度指数变点分析

李德刚

青海师范大学数学与统计学院，青海 西宁

收稿日期：2024年4月2日；录用日期：2024年4月22日；发布日期：2024年4月30日

摘　要

本文主要介绍了CUSUM方法的均值与方差变点估计且该方法对数据的分布限制较少，模拟了CUSUM方法对服从正态分布的时间序列数据以及非正态分布时间序列数据的变点估计，该方法都能准确估计变点的位置，且与实际相符合。对于收集到的“新冠”百度指数时间序列(只存在一个变点)数据应用CUSUM的均值和方差变点估计方法进行变点估计，结果表明均值变点和方差估计方法对“新冠”百度指数的时间序列数据的变点估计都是准确的。然而对于“房贷利率”百度指数时间序列数据(存在两个变点)，需要结合CUSUM方法的递归算法进行变点位置估计；估计到的变点位置与实际位置是相符的。同时对于估计到百度指数时间序列数据变点都具有很好的可解释性，都是符合实际情况的。

关键词

CUSUM方法，变点估计，百度指数

Baidu Index Change Point Analysis Based on CUSUM Method

Degang Li

College of Mathematics and Statistics, Qinghai Normal University, Xining Qinghai

Received: Apr. 2nd, 2024; accepted: Apr. 22nd, 2024; published: Apr. 30th, 2024

Abstract

This paper mainly introduces the CUSUM method's mean and variance change point estimation, which has fewer restrictions on the distribution of data. The CUSUM method's change point estimation for time series data subject to normal distribution and non-normal distribution time series data is simulated. The method can accurately estimate the position of change points and is consistent with reality. For the collected time series data of the "new crown" Baidu index (only one va-

文章引用: 李德刚. 基于 CUSUM 方法的百度指数变点分析[J]. 统计学与应用, 2024, 13(2): 521-527.
DOI: 10.12677/sa.2024.132052

riable point exists), CUSUM's mean and variance variable point estimation methods are applied to estimate the variable points, and the results show that both the mean and variance estimation methods are accurate for the time series data of the "new crown" Baidu index. The results show that the mean and variance estimation methods are accurate for the time series data of "new crown" Baidu index. However, for "mortgage interest rate" Baidu index time series data (there are two change points), it is necessary to combine the recursive algorithm of the CUSUM method to estimate the change point location; The estimated position of the change point is consistent with the actual position. At the same time, it has good interpretability for estimating the change points of Baidu index time series data, and it is in line with the actual situation.

Keywords

CUSUM Method, Change Point Estimation, Baidu Index

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

变点研究在时间序列分析中是一个重要的课题。变点是指时间序列数据在某个时间点发生突然变化的点。变点理论是统计学中的一个经典分支，其基本定义是在一个序列或过程中，当某个统计特性(分布类型、分布参数)在某时间点受系统性因素而非偶然性因素影响发生变化，我们就称该时间点为变点。变点问题最早由 Page [1]于 1954 年应用于质量工程管理，后来逐渐发展到在多个领域都有相关应用，如经济、气象、医疗诊断等领域；详见综述性文献，张学新[2]。积累和方法(CUSUM)是由 Page 提出的，在估计变点方面有着较好的应用，Hawkins [3]使用极大似然估计正态均值变点；Kim H G [4]基于似然比推断线性回归模型变点；Wang [5]利用小波分析估计方差变点；Lee S [6]通过 CUSUM 估计时间序列方差变点；韩四儿[7]研究 ARCH 和 GARCH 模型变点问题。百度搜索引擎是全球最大的中文搜索引擎[8]，每天要响应数十亿次的搜索请求。百度指数是以百度搜索引擎提供的海量行为数据为基础，推出的大数据统计分析平台[9]。自从 2006 年发布以来，百度搜索引擎已经成为很多机构进行大数据分析的重要依据[10]，也是很多学术界研究的重要工具。本文研究百度指数时间序列数据的变点估计问题，使用 CUSUM 方法中的均值和方差变点方法进行准确的变点估计。对百度指数进行变点估计可以更好地认识人们对检索词背后内容的关注度并对变点发生的位置能够有一个很好的可解释性。

2. 模型建立与检验统计量

本文基于 CUSUM 方法的均值与方差变点估计。CUSUM 方法的主要优势在于对序列的分布限制较少，在数据不符合正态分布时可用此方法对序列进行变点分析[11]。所以选择本文 CUSUM 方法对目标序列进行变点估计。首先介绍 CUSUM 方法的均值变点估计；假设 $X_1, X_2, X_3, \cdots, X_n$ 是一组随机变量序列，序列 $\{X_t\}$ 可表示为 $X_t = u + \delta_t + \varepsilon_t$。其中 u 为序列均值 ε_t 为一个均值为 0，方差为 σ^2 的随机误差项，若在 k 时刻出现均值变点，且均值跳跃度为 $\Delta = u_1 - u_2$， u_1 为 k 时刻前的均值， u_2 为 k 时刻后的均值。则有

$$\delta_t = \begin{cases} 0, 1 \le t \le k \\ \Delta, k+1 \le t \le n \end{cases}$$

此时 CUSUM 统计量如下：

$$CUSUM_x(k)=\frac{1}{\sqrt{n}}\left(\sum_{t=1}^{k}X_t-\frac{k}{n}\sum_{t=1}^{n}X_t\right)$$
$$=\frac{k}{n}\left(1-\frac{k}{n}\right)\left[\sqrt{n}\left(\bar{X}_k-\bar{X}_k^*\right)\right]$$
$$=\frac{k(n-k)}{n^2}\left[\sqrt{n}\left(\bar{X}_k-\bar{X}_k^*\right)\right]$$

其中 $\bar{X}_k=\frac{1}{k}\sum_{t=1}^{k}X_t$，$\bar{X}_k^*=\frac{1}{n-k}\sum_{t=k+1}^{n}X_t$。

当统计量 $CUSUM_x(k)$ 在 $1\le k\le n$ 时取得最大值时，我们就能找到均值变化时刻 k 的估计位置 $\hat{k}$。即：

$$\max_{1\le k\le n}\left|CUSUM_x(k)\right|$$
$$\hat{k}=\arg\max_{k}\left|CUSUM_x(k)\right|$$

其次对于均值 u 已知且方差变化时刻为 k_1 情形下的 CUSUM 方法的方差变点估计的 CUSUM 统计量 $V(k)$ 如下：

$$V(k)=\frac{k(n-k)}{k^2}\left[\frac{1}{k}\sum_{t=1}^{k}(X_t-u)^2-\frac{1}{n-k}\sum_{t=k+1}^{n}(X_t-u)^2\right]$$
$$\max_{1\le k\le n}\left|V(k)\right|$$
$$\hat{k}_1=\arg\max_{k}\left|V(k)\right|$$

当 CUSUM 统计量 $V(k)$ 在 $1\le k\le n$ 取得最大值时，我们就能找到方差变化时刻 k_1 的估计位置 $\hat{k}_1$。

从上述对均值和方差变点的估计量的描述可以得出均值变点的估计量形式较简单，且不会被方差变点的影响；而对于方差变点的估计量会受均值的影响，对于估计的均值变点和方差变点分别于各自真实值之间的收敛速度与一致性的研究，详见袁芳[12]。从收敛速度上来看，变点估计与变点前后的跳跃度有关，跳跃度越大，变点估计的精确度就越大。

3. 数值模拟

在数值模拟部分我们主要展示 CUSUM 方法的均值变点估计对服从正态分布的数据与非正态分布的数据进行变点估计；对于方差变点估计的过程与均值变点估计类似。首先随机生成 100 个标准正态分布的时间序列数据，变点位置设定在时间点 25 处并将该点数据增加 3。用均值变点的估计方法对该时间序列数据进行变点估计，结果表明均值变点的估计方法能够准确估计到该变点。输出结果见图 1。

图 1 中我们可以看出均值变点的估计方法估计出了该变点的位置；变点前的均值在−0.03 左右，而变点后的均值在 0.12 左右，图上两条红线分别代表变点前的均值与变点后的均值。可以得到 CUSUM 均值变点方法对服从正态分布的时间序列数据的变点估计结果准确，估计出的变点所在的时间点与实际变点所在的时间点相符合。

随机生成前 100 个数据是服从标准正态分布，后 85 个数据是服从泊松分布的时间序列数据；将生成的数据拼接成一个 185 个数据的时间序列数据，该数据明显是不服从正态分布。变点位置设定在时间点 25 处并将该点数据增加 3。同样采用 CUSUM 均值变点的方法对该时间序列数据进行变点估计。输出结果见图 2。

图 2 可以得用 CUSUM 方法的均值变点估计对于非正态分布的时间序列数据的变点位置估计也是准确的。估计出的变点与实际存在的变点所在时间点一致。综上所述可以得到对于 CUSUM 方法进行变点估计对数据的分布没有服从正态分布的要求。CUSUM 方差变点估计方法的数值模拟过程与 CUSUM 均

值变点方法类似，结果表面都能估计出服从正态分布与非正态分布的时间序列数据变点的位置，且与实际变点的位置相符合。

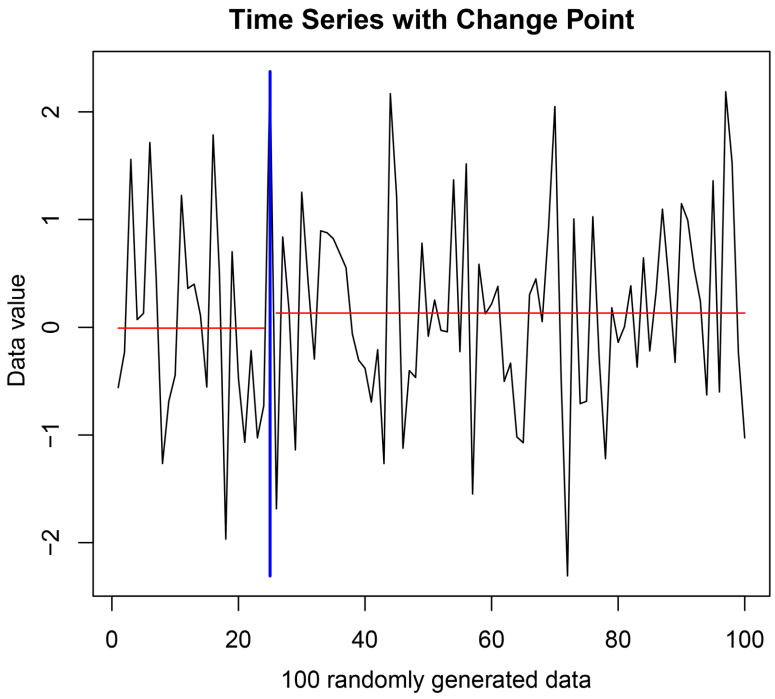

Figure 1. Mean change point estimation results for the standard normal distribution
图 1. 标准正态分布的均值变点估计结果

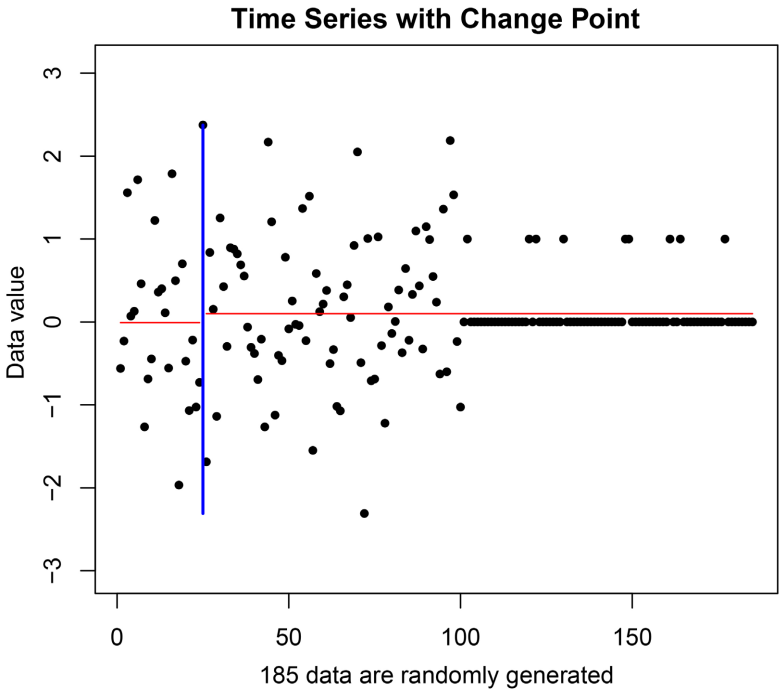

Figure 2. Estimate of mean change point for non-normal distribution data
图 2. 非正态分布数据的均值变点估计

4. 实证分析

百度指数上的搜索包含 PC 端、移动端以及 PC + 移动端三个搜索指数的统计。本次收集的数据来源于 PC + 移动端，分别收集了 2023 年 7 月 22 日~2023 年 10 月 19 日(90 天)的新冠和房贷利率的百度指数时间序列数据。收集到的数据详见表 1 和表 2。

Table 1. 90 days of “New Crown” Baidu index data
表 1. 90 天的“新冠”百度指数数据

2279	2349	3061	3406	3182	3130	2976	2389	2436	2983	2971
3136	3519	3327	2879	3099	3914	3816	4265	5482	6174	5473
4957	5125	5121	5334	6539	5493	5077	5903	7139	5406	4896
4820	4177	3542	3316	3776	3830	3496	3469	2960	2527	2677
3299	3138	3270	3156	3103	2786	2932	3573	3128	3010	3147
3367	2606	2774	3222	3123	3152	3244	3464	2505	2559	2995
3023	2860	2394	1925	1796	1693	1934	1989	2102	2169	2380
2738	2595	2601	6642	19,270	8055	4409	2869	3113	3800	3046
2814	2681									

Table 2. Baidu index data of “mortgage interest rate” for 90 days
表 2. 90 天的“房贷利率”百度指数数据

1525	1375	2021	2250	2167	2211	2825	1669	1552	2238	2252
3845	2888	2504	1694	1506	2231	2072	2086	1995	1760	1302
1289	2032	2806	2926	2479	2671	1626	1853	3402	2240	2088
1676	1797	1306	1198	1693	2297	2523	5751	16,163	8023	4712
5504	5001	3153	10,380	5723	3333	3631	3082	3724	3381	2856
4027	3442	2779	2750	2819	2740	2539	3385	2092	2900	19,496
8079	3826	2684	1475	1351	1349	1278	1206	1261	1214	1282
1991	1917	1919	1842	1739	1648	1525	1150	1163	1698	1594
2002	1484									

画出表 1 和表 2 中的时间序列数据的时间序列图，见图 3、图 4。

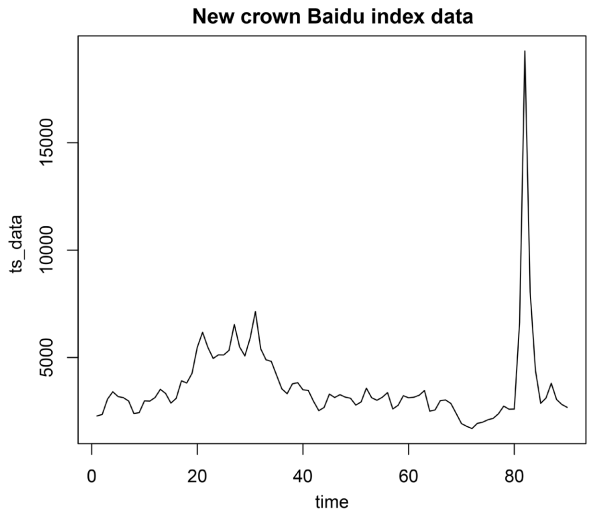

Figure 3. 90-day Baidu index time series of “new crown”
图 3. 90 天的“新冠”百度指数时间序列图

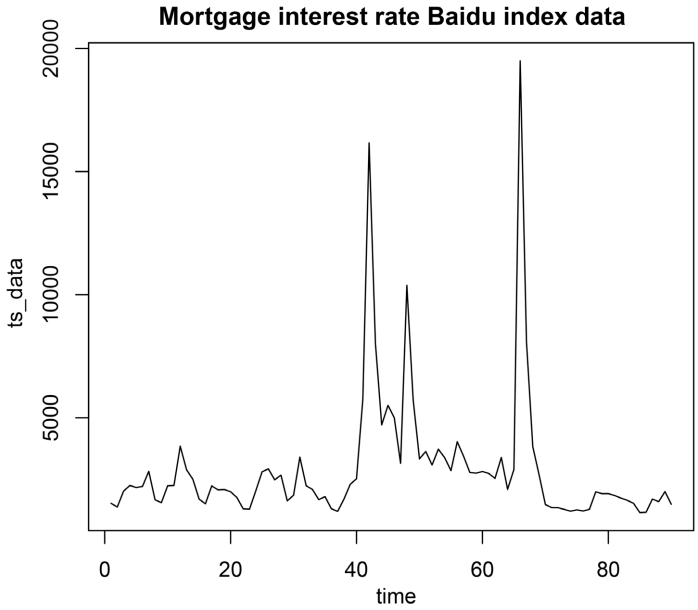

Figure 4. Baidu index data of “mortgage interest rate” for 90 days
图 4. 90 天的“房贷利率”百度指数数据

对于表 1 中的数据进行 Shapiro-Wilk 检验，得到的 p 值远远小于显著性水平 $\alpha = 0.05$，所以拒绝数据服从正态分布的原假设。虽然数据不服从正态分布，但是 CUSUM 方法对不服从正态分布的时间序列数据也能进行变点估计。分别对“新冠”百度指数数据进行基于 CUSUM 方法的均值和方差变点估计；估计结果显示两种方法都能准确的估计出实际变点所在的位置，即时间点 2023 年 10 月 11 日。变点前均值为 3484.198；变点后均值(包含变点)为 5561.889；变点前的方差为 1,454,681，变点后的方差(包含变点) 29,248,767；从均值和方差在变点前后的变化，可以确定该时间点就是变点。疫情结束后人们对“新冠”这个词的检索明显下降很多，关注也较少；所以在变点前该检索词的百度指数较低；但是在变点位置的该检索词的百度指数明显陡增，变点后的百度指数又回到一个较低的水平。具体原因在于发生变点的当天相关媒体在百度上发布了一条关于截至北京时间 2023 年 10 月 11 日全球累计确诊病例超 6.96 亿的新闻，人们在当天对“新冠”的关注度就大幅提升，所以当天关于“新冠”检索词的百度指数很高，所以对变点的估计很准确。

同样对于表 2 的数据做 Shapiro-Wilk 检验，得到的 p 值远小于显著性水平 $\alpha = 0.05$，所以拒绝数据服从正态分布的原假设。采用基于 CUSUM 方法的均值和方差变点估计，该方法能估计对数据不服从正态分布的时间序列数据的变点。从图 4 可以看出变点的位置不止一个，然后 CUSUM 方法只会给出最大统计量值对应的时间点作为变点位置的估计，所以在这里使用递归算法进行变点位置的估计，再估计到第一个变点位置后，将该位置所对应位置的数据替换为整个数据的均值(方差)，然后再次进行基于 CUSUM 方法的变点估计得到第二个变点的位置。结果表明使用递归 CUSUM 均值变点方法对“房贷利率”百度指数数据进行估计时，得到的结果是估计出的两个变点位置都是符合实际的，第一个变点时间点为 2023 年 9 月 1 日，第二个变点的时间点为 2023 年 9 月 25 日；在第一个变点前的均值为 2185.878，第一个变点到第二个变点前的均值为 4505.792，第二个变点后的均值为 2646.92。同时，方差变点的估计方法得到的变点的估计位置也是同变点实际存在位置相符合。第一个变点出现的原因在于国家相关部门在 2023 年 9 月 1 日发布了调整房贷利率的文件政策，该政策具体的实施时间为 2023 年 9 月 25 日，该时间点正是估计到的第二个变点的时间。在后疫情时代的房贷利率成为了人们关心的一个问题，为了缓解

房贷压力，国家相关部门出台了政策对房贷进行降息，即下调了房贷利率，这个政策的出台直接导致了人们对房贷利率更高的关注度，导致人们在这个两个变点的时间增加对“房贷利率”检索词的检索。贷款买房是大部分人在购房时的一个选择，所以在房贷利率出现下调的政策时(即第一个变点的时间点)，房贷利率的百度指数较高，关注的人较多，该政策具体实施的当天(即第二个变点是时间点)，房贷利率的百度指数比政策颁布当天更高；这也说明人们更加关心的是政策具体实施的结果对自我切身利益的影响。

5. 总结与反思

本文主要采用基于 CUSUM 方法的均值与方差变点估计，对“新冠”与“房贷利率”的百度指数时间序列数据进行了变点估计。在对“新冠”百度指数时间序列数据进行变点估计时，均值和方差变点估计方法都准确的估计出了时间序列数据中变点位置，然而对于“房贷利率”的百度指数时间序列数据进行变点估计时，由于 CUSUM 方法一次只能估计出一个变点位置，所以使用了递归算法对多个变点位置进行估计。均值变点与方差变点估计方法都准确估计出了两个变点所在的时间点位置，同时对于估计到时间序列数据变点具有很好的可解释性。对于该方法用于百度指数的变点分析应用具有很好的实际意义；在当前信息化的时代下，信息的传播迅速，可以根据相关信息变点的出现来对下一步的计划做出提前的规划，做好相关的准备，可以给决策者一定的时间去更好的把握变化，做出较优的决定。同时，变点作为生活中的一种突变的情况存在，它的出现都会引起一些意想不到的变化，所以研究变点是十分有必要的。未来对于存在多个变点位置的时间序列的百度指数数据可以尝试在 CUSUM 方法的基础上使用二值分割算法估计变点位置。

参考文献

[1] Page, E.S. (1954) Continuous Inspection Schemes. *Biometrika*, **41**, 100-115. https://doi.org/10.1093/biomet/41.1-2.100

[2] 张学新. 变点估计问题最新进展综述[J]. 江汉大学学报: 自然科学版, 2012, 40(2): 18-24.

[3] Hawkins, D.M. (1977) Testing a Sequence of Observations for a Shift in Location. *Journal of the American Statistical Association*, **72**,180-186. https://doi.org/10.1080/01621459.1977.10479935

[4] Kim, H.J. and Siegmund, D. (1989) The Likelihood Ratio Test for a Change-Point in Simple Linear Regression. *Biometrika*, **76**, 409-423. https://doi.org/10.1093/biomet/76.3.409

[5] Wang, Y. (1995) Jump and Sharp Cusp Detection by Wavelvets. *Biometrika*, **82**, 385-397. https://doi.org/10.1093/biomet/82.2.385

[6] Lee, S. and Park, S. (2001) The CUSUM of Squares Test for Scale Changes in Infinite Order Moving Average Processes. *Scandinavian Journal of Statistics*, **28**, 625-644. https://doi.org/10.1111/1467-9469.00259

[7] 韩四儿. ARCH 模型和 GARCH 模型的变点检测[D]: [硕士学位论文]. 西安: 西北工业大学, 2004.

[8] 张令涛, 赵林, 张亮, 等. 智能电网调控中心变电站图形数据即插即用技术[J]. 电力系统保护与控制, 2018, 46(19): 74-80.

[9] 袁小艳, 贺建英, 成淑萍. 百度指数下数据分析人才需求研究[J]. 电子设计工程, 2023, 31(13): 22-26+31.

[10] 陆晓芬, 亓学成, 王薇, 等. 配网端多能互补工程全寿命周期经济评价方法[J]. 沈阳工业大学学报, 2020, 42(3): 241-246.

[11] 方媛. 气温时间序列的均值变点检验的探究[D]: [硕士学位论文]. 广州: 暨南大学, 2015.

[12] 袁芳, 田铮, 苏晓丽, 等. 独立序列均值与方差变点的累积和估计及应用[J]. 控制理论与应用, 2010, 27(3): 395-399.