Multiple Change-points Test Based on Heteroscedasticity and Its Application in Subway Passenger Flow

X Zhang¹, S Z Shi¹*, H Y Wang²

¹ Science School, Changchun University of Science and technology, China.
² Tianchuang Credit Service Co., Ltd., China.

zhangxue_990603@163.com

Abstract. A class of heteroscedastic multiple change-points test method (H-SMUCE) is introduced to test the mean change-point of subway passenger flow data. And the Algorithm steps is proposed. In this paper, since the passenger flow data obeys the normal distribution, this test method has a high accuracy rate. Simulation shows that the accuracy is over 90%. This algorithm for change-points test is applied to the passenger flow of 28 stations in Shanghai Metro Line 1, the positions and times of the change-points are measured, and the law of the change in average passenger flow is analyzed.

Keywords: change-point; heteroscedasticity; local likelihood ratio; H-SMUCE test; passenger flow

1. Introduction

These The change-point test method is a very challenging research topic in many fields, including economics and finance [1], signal tracking[2], medical diagnosis[3].

In the multiple change-points test, Lee[4] proposed an estimation method for the number of change points in the sequence of independent normal distributions with the same variance. Frick, Munk[5] introduced the SMUCE method for multiple change-points problem in exponential family regression. Balakrishnan N][6] proposed a Wilcoxon-Mann-Whitney statistic that does not require any parameter falseness under the exponential distribution to perform a single change-point test for small-life samples. These methods are only suitable for the same variance and are not robust to heteroscedasticity. For the heteroscedastic change-point test, Fang Yuan and Wei Tian[7] used the CUSUM method to study the change-point estimation problem when the mean and variance of the independent sequence have different change-points. The method only considers the estimation problem that there is a change-point in the mean and variance of the independent sequence. F Pein[8] proposed the H-SMUCE change-point test method, which can adapt to the variation of variance, and the method also shows good feasibility and effectiveness in the ion channel problem.

Change-point test also has many applications in traffic. X Wang, Z Meng[9] applied the local comparison change-point algorithm to analyse the traffic flow failure problems. Nowadays, we have really entered the era of big data transportation with the explosive growth of traffic data. So this new situation inspires us to think the traffic flow test problems based on the deep architecture model with large traffic data. In this paper, the H-SMUCE test method is applied to the traffic rheological point test for the first time, and application of the simulation data is proposed to give the control overestimation probability. Conducting a change-point test on the passenger flow of 28 stations on Shanghai Metro Line 1, and the law of the change in average passenger flow is analyzed.
In Section 2, the multiple change-points test model and the corresponding local likelihood ratio statistic is given. The algorithm steps is proposed based on the H-SMUCE method and corresponding flow charts are given in this section. Section 3 is simulation, and the value of the overestimation probability \( \alpha \) are given, which guarantees that the test accuracy of the method is above 90%. In Section 4, this method is used to detect the passenger flow of the 28 stations on the Shanghai Metro Line 1, to obtain the number and location of the change-points of each station and to carry out the corresponding statistical analysis.

2. Multiple change-points test based on H-SMUCE method

2.1. Change-point test model
Set the change-point model to
\[
Y_i = \mu(i/n) + \sigma(i/n)\varepsilon_i, i = 1, \ldots, n
\]
(1)
Where \( \varepsilon_i \) is in the normal model, which mean is 0 and variance is 1. It is assumed that the standard deviation \( \sigma(i/n) \) can only have a change-point at the same position as the mean \( \mu(i/n) \). \( \mu(\cdot) \) and \( \sigma^2(\cdot) \) is a pair of unknown piecewise constant functions. If there are \( K \) change-points \( \tau_1, \ldots, \tau_K \) on the interval \([0,1]\), then
\[
\begin{align*}
\mu &= \sum_{k=0}^{K} m_k 1_{(\tau_k, \tau_k]}, K \in \mathbb{N}; \\
\sigma^2 &= \sum_{k=0}^{K} s_k^2 1_{(\tau_k, \tau_k]}, K \in \mathbb{N} \\
S &= \{ (\mu(t), \sigma^2(t)) : [0,1] \mapsto \mathbb{R} \times \mathbb{R} \}
\end{align*}
\]
H Li [10] and others regard the variance as constant, and only seek the change in the mean. In this paper, during the change-point test, the variance is regarded as an annoying parameter, and only the change of the mean is considered.

2.2. Test Statistics
Take the observation value \( Y = (Y_1, \ldots, Y_n) \), the model overall obeys the normal distribution, and the local likelihood statistic \( T^i_j \) satisfies
\[
T^i_j(Y, \mu([i/n, j/n])) = (j-i+1) \left( \frac{\overline{Y}_j - \mu([i/n, j/n])}{\hat{s}_y^2} \right) ^2
\]
(3)
Where \( \overline{Y}_j = \frac{1}{(j-i+1)} \sum_{i=1}^{j} Y_i \) is the average of the observations and \( \hat{s}_y^2 = \frac{1}{(j-i)} \sum_{i=1}^{j} (Y_i - \overline{Y}_j)^2 \) is the local variance estimate. The \( K \) is the number of jumps, and the positions \( \tau_1, \ldots, \tau_K \) are unknown, obviously under the normal model \( T^i_j \sim F(1, j-i) \). F Pein[8] proposed the H-SMUCE change-point test method, which uses the test statistic given in formula (4) to test, and appropriately control the overestimation and underestimation probability. Its maximum likelihood statistic is
\[
\arg \min_{\mu(M)} \left[ l(\mu) s.t. \left\{ \max_{\mu(M)} \left\{ T^i_j(Y, \mu([i/n, j/n])) - q_0 \right\} \leq 0 \right\} \right]
\]
(4)
Where \( M \) is a subset of the set \( S \) of all piecewise mean functions, \( |l(\mu)| \) is the number of change-points, \( q_0 \) is the local threshold of the local likelihood statistic. The function value \( \mu([i/n, j/n]) \) is the average of the observations over the corresponding interval\([i/n, j/n]\).

2.3. Algorithm for change-point test
This section gives the algorithm steps based on the H-SMUCE change-point test. Traffic flow data itself has normality, but the data jumps greatly. Therefore, the data is normalized and then the appropriate overestimation probability is found for the sake of convenience. The algorithm steps are as follows:
Step 1: Read the data for normalization. The normalization formula is: 
$$y_i = \frac{y_i - \min\{y_i\}}{\max\{y_i\} - \min\{y_i\}}.$$ 
The Shapiro-Wilk normality test method is used to test the normalized data first, and then the piecewise normality is test;

Step 2: Determine the appropriate overestimation probability $\alpha$ by using simulation data similar to the traffic data, control the estimation accuracy, and use the length of the data and the $\alpha$ and weight to select the threshold $q_k$ (F Pein [10]):
$$q_k \leq \frac{8\log(n)}{2\alpha}.$$

Step 3: Use the local likelihood ratio statistic $T_k \leq$ threshold $q_k$ to determine whether there is a change-point;

Step 4: The maximum likelihood statistic in formula (4), the 95% confidence interval of the change-point position $\tau_k$ and the confidence band of the estimator $2\sigma$ are calculated by dynamic program. The confidence interval and the confidence band are calculated by the variable jumpint and Confband (R package) implementation.

Algorithm flow chart is as follows (see Figure 1):

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{algorithm_flowchart.png}
\caption{Algorithm flow chart.}
\end{figure}

3. Simulation

This section mainly illustrates the data simulated by the mean variance range of real data to find a suitable overestimation probability for application to the practical data.

According to the collected normalized practical data, the variance of the simulated data is determined to be $(0, 0.35)$. For the case of this small variance, 360 normal distribution data with 8 different means and different variances are simulated, and the data amounts are 30, 50, 50, 50, 50, 50, 50, 30. From left to right, the mean values are 0.5, 0.3, 0.9, 0.3, 0.05, 0.3, 0.5, 0.1, and the variances are 0.18, 0.1, 0.15, 0.2, 0.08, 0.11, 0.3, 0.1, respectively. Take $\alpha = 0.1$ and $\alpha = 0.45$ to perform the heteroscedasticity change-point test. The test results are shown in Figure 2.
Figure 2. Simulates the heteroscedastic multiple change-points test of the $\alpha = 0.1$ and $\alpha = 0.45$, the solid line is the estimated change-point, and the dotted line is the confidence band.

Since the number and location of the change-points are known in the simulation data, the test method loses a change-point when $\alpha = 0.1$ for the passenger flow simulation data with small variance, and the fitting effect of the confidence band is not good as shown in Figure 2. When $\alpha = 0.45$, the method has a good test ability, the corresponding confidence interval is significantly reduced, and the confidence band is also well fitted. After 200 simulation tests, the test accuracy is determined to be over 90%.

We use repeated simulations to determine the range $[0.1, 0.99]$ of overestimation probability $\alpha$. And through the simulation to see that when $\alpha$ is small, it will affect the accuracy of the test of the change-point, so in the following practical applications, in order to avoid the loss of the change-point, $\alpha = 0.45$ should be taken.

4. Application of change-point Test in Subway Passenger Flow

The selected data comes from the Shanghai Metro Line 1 on April 1, 2015, which contains 28 stations, the actual passenger flow data of the inbound and outbound gates, and data samples are collected every 5 minutes. The line operates for about 17 hours a day (operating time is 05:30-22:30). There are 18 ordinary stations (non-transfer stations) and 10 interchange stations. The details of the line the number of change-points entering each station are shown in Figure 3.

Figure 3. Line details and to the number of points change of each station enters the station.

In this paper, we test the heteroscedastic multiple change-points of the passenger flow at the entrance and exit of 28 stations on Shanghai Metro Line 1. In order to balance the overestimation and underestimation control, we control the overestimation probability, analyze the change-points of each station and the correlation of the change-point between the stations. Here is an example of the People’s Square Station with transfer station (see Figure 4 and Table 1).
It can be seen that the entrance and exit of People’s Square Station ushered in peak traffic at 7:20-11:30 and 6:40-9:35, 16:10-19:25 and 17:30-19:25, respectively. Other times are basically in the state of flat traffic. And according to the peaks that enter the station in the morning and exit at night, this part of the passenger is a citizen living near the station. Then according to the morning exit and the peak of the station at night (the convex effect is very obvious), it can be analyzed that there may be highly developed commercial and office areas nearby.

5. Conclusions
Firstly, this paper tests the normality of passenger flow data of Shanghai Metro Line 1. A lot of simulations were done, using the H-SMUCE change-point test method, which were then applied to the traffic flow data. The simulation shows that the accuracy of the proposed algorithm reaches over 90%. In the test of the heteroscedastic change-points of the practical data, the change-point was studied and analyzed, the morning peak was 7:00-9:30 and the evening peak was 16:30-19:30 of Shanghai Metro. At the same time, the method also has an abnormality test capability, which enhances its practicability further.

The H-SMUCE change-point test method in this paper can be quickly calculated in real time. In the big data era of transportation, there is a good prospect, and it also provides important data basis for passenger flow guidance and operation scheduling of urban orbit.

Acknowledgments
This work was supported in part the National Natural Science Foundation of China 11601039 and the Jilin Province Natural Science Foundation 20140101199JC.

References
[1] Chen, J. and Gupta, A.K. Parametric Statistical change-point Analysis. Birkhauser, Boston[M].2000 :1-4757.

[2] Fan Wu, Gaojun Xiong, Zhiwei Ye. Application of Wavelet Transform in Signal Mutation Detection[J]. Computer and Modernization, 2008(8):133-135

[3] Chen J , Wang Y P . A statistical change-point model approach for the detection of DNA copy number variations in array CGH data[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2009, 6(4):529-541

[4] Lee CB. Estimating the number of change-points in a sequence of independent normal random variables[J]. Statistics and Probability Letters, 1995, 25:241-248

[5] K Frick, A Munk. Multiscale change-point inference Journal of the Royal Statistical Society, 2014, 76 (3) :495–580

[6] Balakrishnan N , Bordes L , Paroissin C. Single change-point detection methods for small lifetime samples[J]. Metrika, 2016, 79(5):531-551

[7] Fang Yuan, Wei Tian, Xiaoli Su. Accumulation and estimation of independent sequence mean and variance change-points and its application[J]. Control Theory & Applications, 2010, 27(3):395-399

[8] F Pein, H Sieling , A Munk. Heterogeneous change-point inference[J]. Journal of the Royal Statistical Society, 2017,7(4) : 957-1292

[9] X Wang , Z Meng. Study of the local-comparison change-point algorithm to analyze traffic flow breakdown[J]. IEEE International Conference on Systems, 2004, 7 :6186-6191

[10] H Li, A Munk .FDR-Control in Multiscale Change-point Segmentation[J]. Statistics, 2017, 10 (1)