Testing change points and its application in urban transportation

To cite this article: H Y Wang et al 2018 J. Phys.: Conf. Ser. 1039 012045

View the article online for updates and enhancements.
Testing change points and its application in urban transportation

H Y Wang¹, S ZShi¹*, Z X Jiang¹ and H F Jia²

¹ Science School, Changchun University of Science and technology. Changchun, China
² School of Transportation, JiLin University. Changchun, China

Email: whywanghongyu@yeah.net

Abstract. In this paper, we propose the method of testing change points with the exponential family distribution. Under the condition of the same variance, we test whether exist change point in the mean value, we give the test statistics under some conditions when the distribution is unknown, and prove that the asymptotic distribution is the extreme value distribution. The result is same as many literatures. This method is extended to testing multiple change points. By simulating Gamma distribution, Weibull distribution, Log-normal distribution, two-parameter exponential distribution and their mixture distributions, we found the appropriate segment interval size and give the algorithm of the single change point and the multiple change points. We applied this method to testing change points with the pedestrian flow data in Shanghai Railway Station.

Key Words: Change point; Exponential family distribution; Mean test; Asymptotic distribution; Extreme value distribution; Transportation; Simulation Theory

1. Introduction

Studies of change-point detection date back to 1950s. In the past half century, the topic has attracted a great deal of attention in such fields as statistics, engineering, economics and bioscience. For example Miao (1988) proposed the inference in a model with at most one slope-change point; Hsu (1979) proposed the method of detecting shifts of parameter in Gamma sequences with applications to stock price and air traffic flow; Cobb (1978) applied the change point model to hydrology; Braun et al. (2000) solved DNA sequence segmentation by using the multiple change point fitting via quasilikelihood. Fong et al. (2015) employed the change point model-based test in the study of immune responses that are associated with the risk of mother to child transmission of HIV-1.

Depending on the goal of the data that are collected, detection of change points can be crucial for decision making or necessary for understanding certain scientific issues. Many frequentist’s and Bayesian methods have been introduced to detect the change of the mean, variance, slope of regression line, hazard rate, or nonparametric distribution for various models. We refer to the books by Chen (1988); Brodsky and Darkhovsky (1993); Csörgó and Horváth (1997); Chen and Gupta (2011) for various aspects for classical change-point analysis, and to an article Lee (2010) for a list of comprehensive bibliography of books and research papers on this topic.

While this body of work constitutes a rich literature, it mainly deals with the inference of change points when the distributions are known. In fact, the distribution of sequence cannot be judged in some applications. This resulted in some sequence, we couldn’t use these methods of change point test to analyze, such as traffic flow, pedestrian traffic problems etc. But for detecting traffic abnormal incident, analysis and research for the phenomenon of the change of traffic flow is very important. So
regardless of the distribution function, we hope to find a quick way to check the number of change points and their locations. Then analysis the reason. So in this article, we consider the change point test problem when the distribution is unknown and find the minimum data size of the change point by some known distribution, such as Gamma distribution, Weibull distribution, Log-normal distribution, two-parameter Exponential distribution and their mixture distribution. Then quickly find the change point according to the minimum data size.

Let $X_1,\ldots, X_n$ be a sequence of independent random variables with a change point $\tau$, $\tau = k/n$. The distribution function of $X_1, X_2, \ldots, X_k$ is $F(x; \mu_1, \sigma^2)$. The distribution function of $X_{k+1}, X_{k+2}, \ldots, X_n$ is $F(x; \mu_2, \sigma^2)$, $\mu_1, \mu_2, \sigma^2$ are the mean and variance, while $\mu_1, \mu_2, \sigma^2$ are all unknown.

In section 2, we intend to test the location of change points by using the mean test. We give the statistic of mean test. At the same time, asymptotic distribution of estimator of change point is present when the non-negative random variables $X_i$ is a common exponential family distribution. Furthermore, the estimator of change point $\hat{\tau}$ is also present. Section 3 focuses on the simulation results of segmentation of the multiple change point. By simulating the change point test of the common exponential family distribution function to find the minimum data which suitable the change point test. As a practical application of change point problem, in section 4, we research the pedestrian flow at subway gates by change point analysis methods in the common exponential family distribution.

2. The test of change point

Let $X_1,\ldots, X_n$ be a sequence of independent random variables with a common exponential family distribution. The value of mean $\mu$, variance $\sigma^2$ are assumed unknown. The hypotheses can be described in more formal terms as

$$H_0: \mu_i = \mu$$

$(2.1)$

Where the mean of $X_1, X_2, \ldots, X_k$ is $\mu_1$ and the mean of $X_{k+1}, X_{k+2}, \ldots, X_n$ is $\mu_2$. $k$ is unknown. For convenience, we use some notations:

$$\bar{X}_1 = \frac{1}{k} \sum_{i=1}^{k} X_i , \quad \bar{X}_2 = \frac{1}{n-k} \sum_{i=k+1}^{n} X_i , \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i , \quad S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 .$$

Denoting that $\bar{X}_1$ and $\bar{X}_2$ is the estimator of $\mu_1$ and $\mu_2$, respectively. $S^2$ is the estimator of $\sigma^2$. The statistic $T_k$ defined in (2.2) can test whether there was a change point along the sequence.

$$T_k = \sqrt{\frac{k(n-k)}{n}} \left( \frac{1}{k} \sum_{i=1}^{k} X_i \frac{1}{n-k} \sum_{i=k+1}^{n} X_i \right)$$

$(2.2)$

The following theorem give the asymptotic distribution of $\max_{1\leq k\leq n} |T_k|$, we give the lemma of the domain of attraction of the normal law shown in appendix and the following theorem.

**Theorem 2.1** Let the random variables $X_1,\ldots, X_n$ be a sequence of non-degenerate and nonnegative i.i.d. random variables. The distribution of $X_t$ is an exponential family distribution, $Ee^{xX_t} < +\infty$, $0 \leq t \leq 1$, the four order moments of $X_t$ exists, then

$$\lim_{n \to \infty} p\{\max_{1 \leq k \leq n} |T_k| \leq a(x, \log n)\} = \exp\{-2e^{-x}\} \quad -\infty < x < +\infty$$

$(2.3)$

$a(x, \log n)$ is given for

$$a(x, T) = [2\log T]^{-\frac{1}{2}} \left\{ x + 2\log T + \frac{1}{2}\log \log T - \frac{1}{2}\log \pi \right\}$$

$(2.4)$

The proof of the Theorem 2.1 is shown in the appendix. The method borrows an idea advanced by Tan and Miao (2005). Theorem 2.1 provides a method to test the null-hypothesis $H_0: \mu_1 = \mu_2$. Given the test level $\alpha$, from the equation

$$\exp\{-2e^{-x}\} = 1 - \alpha ,$$
We can get $x(\alpha) = -\log[-\frac{1}{2}\log(1-\alpha)]$ if and only if when $\max_{i \in \mathbb{N}} |T_i| > a(x(\alpha), \log n)$, we refused $H_0$, which means there is a change point exists. If there is a change point in the test results, the change point is estimated to be

$$\hat{k} = \frac{1}{n} \min_{i \in \mathbb{N}} \{k : |T_i| = \max_{j \in \mathbb{N}} |T_j|\}$$

(2.5)

The problem of making inference in this model is important in practical applications and of much theoretical interest. Many authors have contributed to it. To name a few among others, Chen (1988) use this method in Normal distribution.

3. Simulation

When the sequence have multiple change points, we need to segment the sequence. In order to verify the feasibility of the statistics given in Section 2 for testing the location of change points, and find out the smallest sample size of the common variable point test in the exponential family as the segmentation basis of the multiple point test, we give the following simulation.

3.1. Simulation test of a single change point

By using the mean change point test, observation data are searched and tested as follow. First, simulating exponential family distribution. Each distribution is divided into two parts which have different means and same variance. The sample size of two parts is $N_1$ and $N_2$ respectively, total sample size is $N = N_1 + N_2$. Then, we use the mean test proposed in section 2 to find the location of change point $\hat{k}$, and then compared with the actual change point $k_0$. Finally, repeat experiment 500 times, when the probability of $P > 0.9$, output $N$, otherwise, improve the number of samples to continue to repeating the experiment.

In the simulation, we find that when the mean and the variance are more than 10, the probability of finding the change point is about 50%, but in reality, the mean and variance in the field of traffic exceed this number. In order to solve this problem, we normalize the data to carry out the change point test. The normalization method is carried out by the equation (3.1). Kao (2010)

$$X_i - \min_{i \in \mathbb{N}} \{X_i\} \over \max_{i \in \mathbb{N}} |X_i| - \min_{i \in \mathbb{N}} |X_i|$$

(3.1)

According to the simulation results, we find that when the sample size of each sequence is 59, the probability of finding the position of the change point is more than 90%. Therefore, we choose 59 for segmented data. The results of the four distributions (Gamma distribution, Weibull distribution, Log-normal distribution, two-parameter Exponential distribution) of exponential families and their mutual combination are shown in Table 1,

|                  | $\mu_1$ | $\mu_2$ | $\sigma^2$ | $n$ | $k_0$ | $\hat{k}$ | $P$ | $\mu_1$ | $\mu_2$ | $\sigma^2$ | $k_0$ | $\hat{k}$ | $P$ |
|------------------|---------|---------|------------|-----|-------|----------|-----|---------|---------|------------|-----|----------|-----|
| Gamma            | 8       | 4       | 2          | 59  | 20    | 20       | 0.9999 | 12      | 100     | 20         | 30  | 30       | 0.9999 |
| Weibull          | 2.7     | 11      | 2          | 59  | 20    | 20       | 0.9999 | 14.3    | 27.8    | 29.5       | 30  | 30       | 0.9999 |
| Log-normal       | 2       | 5       | 1          | 59  | 20    | 20       | 0.9878 | 30      | 70      | 400        | 30  | 30       | 0.9200 |
| Exponential      | 3       | 6       | 10         | 59  | 20    | 20       | 0.9999 | 20      | 60      | 10         | 30  | 30       | 0.9999 |
| Gamma,Weibull    | 8       | 4       | 2          | 59  | 30    | 30       | 0.9800 | 40      | 20      | 10         | 30  | 30       | 0.9900 |
| Gamma,Log-normal | 10      | 15      | 5          | 59  | 30    | 30       | 0.9600 | 10      | 15      | 25         | 30  | 30       | 0.9900 |
| Gamma,Exponential| 12      | 4       | 4          | 59  | 30    | 30       | 0.9999 | 72      | 26      | 36         | 30  | 30       | 0.9999 |
| Weibull,Log-normal| 6      | 14      | 9          | 59  | 30    | 30       | 0.9999 | 40      | 15      | 10         | 30  | 30       | 0.9999 |
| Weibull,Exponential| 3     | 7       | 2          | 59  | 30    | 30       | 0.9999 | 30      | 15      | 10         | 30  | 30       | 0.9999 |
| Exponential,Log-normal | 2 | 10    | 4          | 59  | 30    | 30       | 0.9999 | 19      | 50      | 81         | 30  | 30       | 0.9999 |
3.2. Simulation test of multiple change points
According to the simulation results in section 3.1, this paper select the change points of
distribution to repeat 150 times the simulation of multiple change points. We divide the sample
into several groups, each group has 59 samples, and test the change point of each group.
When we find the change point, recording the position. Choosing the next 59 data from this
position (not include this data) to find the next position of change points, and so on. When
there is no change point in group n, increase the size of sample to several times of 59, repeat the
steps above until find the change point. The sample size of the next change point analysis is
still 59, and finally find all the change points.

We select 310 data for the Gamma distribution, Weibull distribution, Log-normal distribution,
Two-parameter Exponential distribution and mixed distribution, respectively. The mixed distribution
is the mixture of four distributions. The order is Gamma distribution, Weibull distribution, Log-
normal distribution, Two-parameter Exponential distribution, Gamma distribution, Weibull distribution, Log-
normal distribution and Two-parameter Exponential distribution. They have 8 different means and
same variance. Their mean are 20, 40, 60, 80, 100, 80, 50, 30; Their variance are 5, 1, 1, 4, 1, 1 and 10,
10, 64, 81, 10. The size of data are 30, 30, 50, 20, 40, 50, 50, 50. Their shape are stepped. Through
MATLAB simulation, one of the results is shown in Figure 2

\[\text{Figure 2. Simulation of four Exponential Family distribution and Mixed Distribution by multivariate change point test estimated mean, the variance is 5,1,1,4,1,10, in which the dotted line is the true value and the solid line is the estimated mean line chart. Where the abscissa indicates the position of the change point, the ordinate indicates the size of the change point.}\]

4. Application in transportation
In this section, we select April 1, 2015 Shanghai Metro Line 1 pedestrian flow data. We analyze the
data of gate entrance and exit obtained by every 5 minutes in Shanghai Metro Line 1 pedestrian flow.
The number of pedestrian flow at the entrance of the subway gates is 212 and the number of the exit is
210. Firstly, the data were normalized. Then, the method proposed in Section 2 and the simulating
process in Section 3 were used to test the multiple change points. Finally, the results are shown in
Figure 3.

\[\text{Figure 3. The pedestrian flow of the entrance and exit gates of the Shanghai Railway Station subway, in which the dotted line is the true value and the solid line is the line graph of the estimated mean. Where the abscissa indicates the time at which the change point occurred, the ordinate indicates the size of the change point.}\]

From Figure 3 we can see that 7:30-9:20 and 17:00-19:05 are the peaks of inbound stations, the
means are 337 and 527 respectively; 7:15-9:15 is the peak of outbound stations, the mean is
690 people, a larger flow of people. According to the Shanghai Railway Station timetable we can know that in the morning 7:00-9:30, 20 trains reach the Shanghai Railway Station, and this period of time is the morning peak, we can verify Shanghai is the most interprovincial office workers city as reported in news.

We can get the following conclusions from the results of the application: this method is suitable for the change points whose distance is large enough. The estimation location of change points is basically consistent with the actual situation. The distribution of samples is the common exponential distribution. The size of sample is more than 59. Conclusion

In this paper, we propose a fast testing method to find the change points when the distribution is unknown. The simulation results show that the test results are accurate under certain conditions, but when the number of data between two change points is less than 10, this method is difficult to detect the change points. Much work remains to be done for this problem.

References

[1] Braun J V, Braun R K, and Muller H G. Multiple change point fitting via quasilikelihood, with application to dna sequence segmentation. Biometrika, 87(2):301–314(14), 2000.
[2] Brodsky B E and Darkhovsky B S. Nonparametric Methods in Change-Point Problems. Kluwer Academic Publishers, 1993.
[3] Chen J and Gupta A K. Parametric Statistical Change Point Analysis. Springer Basel AG, 2011.
[4] Chen X R. Inference in a sample change point model. Scientia Sinica, 6:654–667, 1988.
[5] Cobb G W. The problem of the nile: Conditional solution to a change point problem. Biometrika, 65(2): 243-251, 1978.
[6] Csörgő M and Horváth L. 20 nonparametric methods for change point problems. Handbook of Statistics, 7(88):403–425, 1988.
[7] Csörgő M and Horváth L. Limit theorems in change-point analysis. John Wiley and Sons Ltd Chichester, 1997.
[8] Csörgő M, Szyszkowicz B, and Wang Q Y. Donsker’s theorem for self-normalized partial sums processes. Annals of Probability, 31(3):1228-1240, 2003.
[9] Fong Y, Di C., and Permar S. Change point testing in logistic regression models with interaction term. Statistics in Medicine, 34(9):1483-94, 2015.
[10] Giné E, Götze Friedrich, and Mason David M. When is the student t-statistic asymptotically standard normal? Annals of Probability, 25(3):1514-1531, 1997.
[11] Hsu D A. Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis. Journal of the American Statistical Association, 74(365):31-40, 1979.
[12] Kao C. Fuzzy data standardization. IEEE Transactions on Fuzzy Systems, 18(4):745-754, 2010.
[13] Lee T S. Change-point problems: Bibliography and review. Journal of Statistical Theory and Practice, 4(4):643-662, 2010.
[14] Miao B. Q. Inference in a model with at most one slope-change point. Journal of Multivariate Analysis, 27(2):375-391, 1988.
[15] Tan C C and Miao B Q. Inference in γ-distribution with at most one change-point. J. univ. sci. technol. china, (1):51-58, 2005.
[16] Zhu. H L. Statistical inference in the jump change-point estimator. Master’s thesis, Annals of Probability, 2014.

Acknowledgement

This work was supported by NSF grant 51378076 from the National Natural Science Found in China and grant 20140101199JC from the Natural Science Foundation of Jilin Province