World Academy of Science, Engineering and Technology
International Journal of Energy and Power Engineering
Vol:17, No:6, 2023

Change Point Analysis in Average Ozone Layer Temperature Using Exponential Lomax Distribution

Amjad Abdullah, Amjad Yahya, Bushra Aljohani, Amani S. Alghamdi

Open Science Index, Energy and Power Engineering Vol:17, No:6, 2023 publications.waset.org/10013121.pdf

Abstract—Change point detection is an important part of data analysis. The presence of a change point refers to a significant change in the behavior of a time series. In this article, we examine the detection of multiple change points of parameters of the exponential Lomax distribution, which is broad and flexible compared with other distributions while fitting data. We used the Schwarz information criterion and binary segmentation to detect multiple change points in publicly available data on the average temperature in the ozone layer. The change points were successfully located.

Keywords—Binary segmentation, change point, exponential Lomax distribution, information criterion.

I. Introduction

TIME series data have become increasingly relevant in numerous fields, including medicine, modeling, finance, industry, meteorology, and entertainment. They are sequences obtained through measurements over time demonstrating a system's behavior. Those patterns may shift gradually due to external occurrences and/or changes in internal organizational dynamics or distribution (see [1]). Change point detection (CPD) refers to the problem of identifying a time series change due to an abrupt change in data. There are several approaches to conducting change point analysis, including the information approach, the likelihood ratio test, and the Bayesian method. In this paper, we study CPD by applying the Schwarz information criterion (SIC) for the exponential Lomax (EL) distribution to data on the average ozone layer temperature. Our goal is to show that these time series data can serve in the development of new methodologies for CPD.

A. Concept of an Information Approach

According to [2], the change point problem can be defined as follows. Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables with distribution functions $F_1, F_2, .., F_n$, respectively. In general, the change point problem is thus to test the following null hypothesis:

$$H_0 : F_1 = F_2 = ..F_n$$

versus the alternative:

$$H_1 : F_1 = .. = F_{k1} \neq F_{k1+1} = .. = F_{k_q} \neq F_{kq+1} = .. = F_n$$

where $1 < k_1 < k_2 < ... < k_q < n; q$ is the unknown number of change points; and $k_1, k_2, .., k_q$ are the respective unknown positions that need to be estimated. The change point problem is to test the null hypothesis about the population parameter $\theta_i,\ i = 1, .., n$, if the distributions $F_1, F_2, .., F_n$ belong to a common parametric family $F(\theta)$, where $\theta \in R^p$:

$$H_0 : \theta_1 = \theta_2 = .. = \theta_n = \theta \text{ (unknown)},$$

versus the alternative hypothesis:

$$H_1 : \theta_1 = .. = \theta_{k_1} \neq \theta_{k_1+1} = .. = \theta_{k_2} \neq \theta_{k_2+1} = .. = \theta_{k_q-1} \neq \theta_{k_q} = .. = \theta_n$$

where q and $k_1, k_2..., k_q$ need to be estimated. Detecting a change point using these hypotheses involves determining the occurrence of the change point in a dataset and estimating the number and location of change points. The SIC technique is to find the model that minimizes the

$$SIC(k) = -2 \log L\left(\hat{\theta}_k\right) + \dim\left(\hat{\theta}_k\right) \log(n), \text{ for } k = 1, 2, \ldots, K, \tag{1}$$

where n is the sample size and K is the number of the model parameters.

Under the null hypothesis H_0, $SIC(n)$ can be defined as follows:

$$SIC(n) = -2 \log L\left(\hat{\theta}\right) + \dim\left(\hat{\theta}\right) \log(n), \tag{2}$$

Therefore, we do not reject H_0 if:

$$SIC(n) \leq \min_{k_0 \leq k \leq n-k_0} SIC(k), \tag{3}$$

where k_0 is selected such that the maximum likelihood estimation (MLE) can be calculated accurately. We reject H_0 if:

$$SIC(n) > SIC(k), \tag{4}$$

The change point location is determined to be $\hat{k}$, such that:

$$SIC(\hat{k}) = \min_{k_0 \leq k \leq n-k_0} SIC(k) \tag{5}$$

for some k. We note that the SIC's general penalty affects only the dataset and the number of parameters to be estimated. Zhang and Siegmund [3] suggested that the SIC will identify change points more efficiently when the change points are in the center of the data. However, if the change points are located at the beginning or the end of the data, the approach applied in the change point problem may not detect them. For these reasons, we need sufficient observations to compute the parameters, and we calculate the $SIC(k)$ for $k_0 \leq k \leq n-k_0$,

A. Alghamdi is with the Statistics Department, Faculty of Science, King AbdulAziz University, Jeddah, Saudi Arabia (e-mail: amaalghamdi@kau.edu.sa).

Open Science Index, Energy and Power Engineering Vol:17, No:6, 2023 publications.waset.org/10013121.pdf

where k_0 is chosen to be large enough that the MLE can be calculated rigorously.

B. Literature Review

Some statisticians have contributed to solving the change point problem to detect change points, if they exist. Chernof and Zacks [4] studied the Bayes estimator for the current mean of a normal distribution. Worsley [5] studied the identification of a single change point when variance is known and unknown. Kim and White [6] detected a single change point in a simple linear regression model using the likelihood ratio test. Ning and Gupta [7] investigated the change point problem for the generalized lambda distribution. Matteson and James [8] applied multiple change point analysis of multivariate data using none-parametric approach.

Detecting multiple change points is an important challenge that has attracted many researchers, as it can solve real-life problems. Vostrikova [9] suggested using binary segmentation to detect multiple change points and their positions. This process has the benefits of simultaneously detecting more than one change point and the corresponding positions and reducing computational time by a great amount. The method of binary segmentation can be explained as follows.

Step 1. Test the null hypothesis given by:

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_n = \theta(\text{unknown}) \tag{6}$$

versus the alternative:

$$H_1 : \theta_1 = \ldots = \theta_k \neq \theta_{k+1} = \ldots = \theta_n \tag{7}$$

where k is the position of the change point at this step. If the null hypothesis H_0 is not rejected, we conclude that there is no change point, and we stop the detection. However, if we reject H_0, a change point has occurred, and we move on to *the next step*.

*Step 2.*Separately check the two sections of the data before and after the change point detected in Step 1 to identify any other change point, if such change points exist.

Step 3. Repeat the previous steps until we find that there is no change point.

Step 4. All of the change points obtained in the previous steps are denoted by $\left\{\hat{k}_1, \ldots \hat{k}_2, \ldots \hat{k}_q\right\}$ with estimated number of change points q.

The rest of this paper is organized as follows. Section II introduces the EL distribution and the change point problem of its three parameters. Application to a real dataset is explained in Section III. Finally, Section IV presents the results of the study.

II. Exponential Lomax Distribution

The Lomax distribution is a known heavy-tailed distribution that models non-negative data. It is named according to [10] and is also conditionally known as the Pareto Type II distribution. The Lomax distribution can be derived from the generalized Pareto distribution as a special case that is used in a variety of forms throughout the literature. The equilibrium of the Lomax probability distribution and its order statistics was analyzed in [11]. However, in modeling results, this distribution is not flexible in modeling data. Thus, shape, location, or scale parameters can be added to derive a new generalized distribution, which may be more flexible, and subsequently study the behavior of the new proposed distribution. Many generalized distributions have been suggested in the literature. Al-Awadhi and Ghitany [12] presented a Lomax mixing distribution for the Poisson parameters and obtained the discrete Poisson–Lomax distribution. Zubair et al. [13] studied the logistic Lomax distribution and its applications to real data. The EL distribution is another extension of the Lomax distribution that was derived and studied by [14].

Let $G(x)$ denote the cumulative density function (CDF) of the Lomax distribution and $f(t)$ the probability density function (PDF) of the exponential distribution. The CDF for the EL distribution is given by:

$$\begin{aligned} F(x;\alpha,\lambda,\beta) &= \int_0^{\frac{1}{1-G(x;\alpha,\lambda)}} f(t;\beta)dt \\ &= \int_0^{\frac{1}{\left(\frac{\lambda}{x+\lambda}\right)^{\alpha}}} \beta e^{-\beta t}dt, \alpha,\beta,\lambda > 0, \end{aligned} \tag{8}$$

and the corresponding PDF, is given by:

$$f(x) = \frac{\alpha\beta}{\lambda}\left(\frac{\lambda}{x+\lambda}\right)^{-\alpha+1} e^{-\beta\left(\frac{\lambda}{x+\lambda}\right)^{-\alpha}}, x \geq -\lambda, \alpha, \lambda, \beta > 0, \tag{9}$$

where α is the shape parameter and λ and β are the scale parameters of the EL distribution.

A. Change Point Problem

Change points are abrupt changes in time series data. Detecting change points is very helpful in time series forecasting and simulations in many fields, such as medical tracking, climate change identification, picture and language processing, and human interaction. The change point problem for the shape and scale parameters of the EL distribution using the SIC can be explained as follows.

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables from the EL distribution with scale parameters λ and β and shape parameter α. The change point is detected by checking the following null hypothesis:

$$H_0 : \left.\begin{array}{c} \alpha_1 = \alpha_2 = \ldots = \alpha_n = \alpha \\ \lambda_1 = \lambda_2 = \ldots = \lambda_n = \lambda \\ \beta_1 = \beta_2 = \ldots = \beta_n = \beta \end{array}\right\} (unknown) \tag{10}$$

versus the alternative:

$$H_1 : \begin{array}{c} \alpha_1 = \ldots = \alpha_k = \alpha^{'} \neq \alpha_{k+1} = \ldots = \alpha_n = \alpha^{''} \\ \lambda_1 = \ldots = \lambda_k = \lambda^{'} \neq \lambda_{k+1} = \ldots = \lambda_n = \lambda^{''} \\ \beta_1 = \ldots = \beta_k = \beta^{'} \neq \beta_{k+1} = \ldots = \beta_n = \beta^{''}, \end{array} \tag{11}$$

where $1 < k < n$ is the unknown location of the change point. Under the null hypothesis, the SIC is defined as:

$$SIC(n) = -2\sum_{i=1}^{n} log(f(x_i;\hat{\alpha},\hat{\lambda},\hat{\beta})) + 3log(n), \tag{12}$$

where $\hat{\lambda}$, $\hat{\beta}$ and $\hat{\alpha}$ are the MLEs of the scale parameters λ and β and the shape parameter α, respectively. Under the alternative hypothesis, SIC is defined as follows:

$$SIC(k) = -2\sum_{i=1}^{k} log(f(x_i; \hat{\alpha}^{'}, \hat{\lambda}^{'}, \hat{\beta}^{'})) - 2\sum_{i=k+1}^{n} log(f(x_i; \hat{\alpha}^{''}, \hat{\lambda}^{''}, \hat{\beta}^{''})) + 6log(n), \quad (13)$$

where $\hat{\alpha}^{'}$, $\hat{\lambda}^{'}$ and $\hat{\beta}^{'}$ and $\hat{\alpha}^{''}$, $\hat{\lambda}^{''}$ and $\hat{\beta}^{''}$ are the MLEs of α, λ and β fitted to the two sections of the data before and after the change point, respectively. Therefore, binary segmentation can be applied to detect multiple change points.

III. Application to Real Data

A. Conversion of DataSet to Independent

In this section, we demonstrate the significance of CPD using the dataset of the average temperature of the ozone layer between January 1, 1998, and December 31, 1998. These data were collected in the Houston, Galveston, and Brazoria areas of Texas and are available at [3]. We consider several change points using the binary segmentation method by applying an SIC testing procedure. The package bbmle in the R statistical software developed by [16] was implemented. We note that the data of the average ozone layer temperature may not be independent. Hence, according to [17], the following data transformation into a separate R_t series was considered as follows:

$$R_t = \frac{P_{t+1} - P_t}{P_t}, \text{ for } t = 1, 2, \ldots, 360 \quad (14)$$

After transforming the data, we tested whether they were independent using the Portmanteau test provided by [18]:

$$Q_k = n\sum_{i=1}^{k} r_i^2 \quad (15)$$

where r_i is the autocorrelation function (ACF) at lag i, and k is the number of lags for which the ACF is considered. Under the null hypothesis of independence, the test statistic has an asymptotic χ^2distribution with degree of freedom k. Using the Portmanteau test, we get:

$$Q_{25} = 360 \times \sum_{i=1}^{25} r_i^2 = 360 \times 0.08411352 = 30.19675 < \chi^2_{0.95}(25) = 37.65248 \quad (16)$$

Therefore, the null hypothesis H_0 is not rejected, and the data are independent. Fig. 1 shows the ACF of the transformed data.

B. Result

To detect change points in the average ozone layer temperature dataset, we applied the test statistics described in (12) and (13) using the EL distribution. We get $SIC(n) = -3.370954 > \min_{2\leq k\leq 358} SIC(k) = SIC(335) = -238.1777$, which shows that a change point is located at position 335 (corresponding to December 6, 1998). Another change point can be observed using binary segmentation by checking the subsequences before and after the change point $\hat{k}$ =335. We found that $SIC(n) = -290.9045 > \min_{2\leq k\leq 333} SIC(k) = SIC(108) = -330.1299$. Therefore, there is a change point at location 108, corresponding to April 18, 1998. The period of January to April 1998 was the warmest of the century, according to the National Climatic Data Center. This occurred due to a phenomenon called sudden stratospheric warming, which is the result of air being pushed in a downward motion in the late winter and spring at high latitudes. Sudden stratospheric warming can considerably alter temperature-dependent chemical reactions of ozone and other reactive gases in the stratosphere and affect the development of features such as ozone holes. Ozone depletion is not limited to the area over the South Pole: Research has revealed that ozone depletion occurs over latitudes that include North America; Europe; Asia; and much of Africa, Australia, and South America (see [19]).

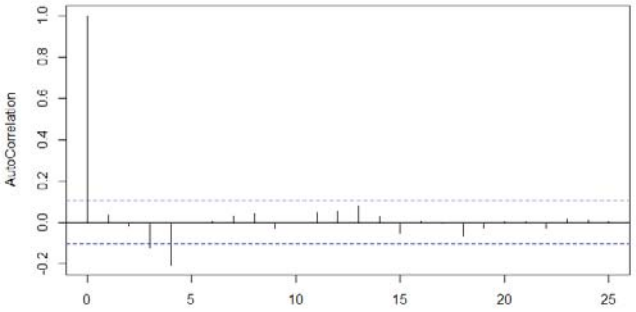

Fig. 1 The autocorrelation function of the transformed Ozone average temperation data

IV. Conclusions

In this article, we studied the change point problem for the EL distribution. We detected multiple change points in the parameters of the proposed distribution by using SIC and applying binary segmentation. Two change points were observed on the average temperature of the ozone layer in 1998. The benefit of CPD in this context is identifying the exact location of the change in temperatures in the ozone layer.

References

[1] Montanez, G. D.; Amizadeh, S.; Laptev, N. Inertial hidden markov models: Modeling change in multivariate time series. AAAI-15. 2015, 10, 1819–1825.
[2] Chen, Jie; Arjun K. Gupta. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*, 2nd ed.; Birkhäuser: Basel, Switzerland, 2011; pp. XIII, 273.
[3] Zhang, N. R.; Siegmund, D. O. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics. 2007, 63 22–32.
[4] Chernof f, H.; Zacks, S. Estimating the current mean of a normal distribution which is subjected to changes in time. Ann. Math. Stat. 1964, 35, 999–1018.
[5] Worsley, K. J. On the likelihood ratio test for a shift in location of normal populations. JASA. 1979, 74, 365–367.

Open Science Index, Energy and Power Engineering Vol:17, No:6, 2023 publications.waset.org/10013121.pdf

[6] Kim, T.-H.; White, H. On more robust estimation of skewness and kurtosis. Finance Res. Lett. 2004, 1, 56–73.
[7] Ning, W.; Gupta, A. K. Change point analysis for generalized lambda distribution. Commun. Stat. Simul. Comput. 2009, 38, 1789–1802.
[8] Matteson, D. S.; James, N. A. A nonparametric approach for multiple change point analysis of multivariate data. JASA. 2014, 109, 334–345.
[9] Vostrikova, L. Detecting "disorder" in multidimensional random processes. Soviet Math. Dokl. 1981, 24, 55–59.
[10] Lomax, K. Business failures: Another example of the analysis of failure data. JASA. 1954, 49, 847–852.
[11] Devi, B.; Kumar, P.; Kour, K. Entropy of lomax probability distribution and its order statistic. IJSS. 2017, 12, 175–181.
[12] Al-Awadhi, S. A.; Ghitany, M. E. Statistical properties of Poisson-Lomax distribution and its application to repeated accidents data. J. Appl. Stat. Sci. 2001, 10, 365–372.
[13] Zubair, M.; Cordeiro, G. M.; Tahir, M. H.; Mahmood, M.; Mansoor, M. A study of logistic-lomax distribution and its applications. J. Prob. Stat. Sci. 2017, 15, 29–46.
[14] El-Bassiouny, A. H.; Abdo, N. F.; Shahen, H. S. Exponential lomax distribution Int. J. Comput. Appl. 2015, 121, 24–29.
[15] Ozone Level Detection Data Set. Available online:https://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection (accessed on 30 July 2021).
[16] bbmle: Tools for general maximum likelihood estimation. Available online:https://rdrr.io/rforge/bbmle/ (accessed on 31 July 2021).
[17] Hsu, D. A. Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis. JASA. 1979, 74, 31–40.
[18] Ngunkeng, G.; Ning, W. Information approach for the change-point detection in the skew normal distribution and its applications Seq. Anal. 2014, 33, 475–490.
[19] Basic Ozone Layer Science. https://www.epa.gov/ozone-layer-protection/basic-ozone-layer-science (accessed on 15 August 2021).

Open Science Index, Energy and Power Engineering Vol:17, No:6, 2023 publications.waset.org/10013121.pdf