Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 832-849
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Consistency of the φ-Divergence Based Change Point Estimator

Mwelu Susan[1], Anthony G. Waititu[2], Peter N. Mwita[3], Charity Wamwea[2]

[1]Pan-African University Institute of Basic Sciences, Technology and Innovation, Nairobi, Kenya
[2]Department of Statistics and Actuarial Sciences, JKUAT, Nairobi, Kenya
[3]Department of Mathematics, Statistics and Actuarial Sciences, Machakos University, Machakos, Kenya
Email: suemwelu@gmail.com

How to cite this paper: Susan, M., Waititu, A.G., Mwita, P.N. and Wamwea, C. (2020) Consistency of the φ-Divergence Based Change Point Estimator. *Open Journal of Statistics*, **10**, 932-849.
https://doi.org/10.4236/ojs.2020.105048

Received: July 20, 2020
Accepted: October 24, 2020
Published: October 27, 2020

Abstract

This paper utilizes a change-point estimator based on the ϕ-divergence. Since we seek a near perfect translation to reality, then locations of parameter change within a finite set of data have to be accounted for since the assumption of stationary model is too restrictive especially for long time series. The estimator is shown to be consistent through asymptotic theory and finally proven through simulations. The estimator is applied to the generalized Pareto distribution to estimate changes in the scale and shape parameters.

Keywords

Change Point, Consistency, φ-Divergence, Kullback-Leibler, Generalized Pareto Distribution

1. Introduction

Let $x_1, \cdots, x_n$ be a time series of size n. Methods in literature consider stationary models in explaining the underlying data generating process. However, stationarity is arguably a very strong assumption in many real-world applications as process characteristics evolve over time. Reviewed literature reveals that the use of one model may not be appropriate to model a non-stationary series and as such various change-point estimation methods have been proposed. However, they are limited in different ways and their suitability depends on the underlying assumptions. Statistical research works have shown that with time, the underlying data generating processes undergo occasional sudden changes [1]. A change point is said to occur when there exists a time $\tau \in \{1, \cdots, n-1\}$ such that the statistical properties of $x_1, \cdots, x_\tau$ and $x_{\tau+1}, \cdots, x_n$ are different. In its simplest

form, change-point detection is the name given to the problem of estimating the point at which the statistical properties of a sequence of observations change [2]. The overall behavior of observations can change over time due to internal systemic changes in distribution dynamics or due to external factors. Time series data entail changes in the dependence structure and therefore modelling non-stationary processes using stationary methods to capture their time-evolving dependence aspects will most likely result in a crude approximation as abrupt changes fail to be accounted for [3]. Each change point is an integer between 1 and $n-1$ inclusive. The process X is assumed to be piece-wise stationary implying that some characteristics of the process change abruptly at unknown points in time. The corresponding segments are then said to be homogeneous within but each of the subsequent segments is heterogeneous in characteristics. For a parametric model the parameters associated with the i^{th} segment denoted θ_i, are assumed to contain changes. Parametric tests for change point are mainly based on the likelihood ratio statistics and estimation based on the maximum likelihood method whose general results can be found in [4].

Detection of change points is critical to statistical inference as a near perfect translation to reality is sought through model selection and parameter estimation. Parametric methods assume models for a given set of empirical data. Within a parametric setting change, points can be attributed to change in the parameters of the underlying data distribution. Generally, change point methods can be compared based on general characteristics and properties such as test size, power of the test or the rate of convergence to estimate the correct number of change point and the change-point locations. Change point problems can be classified as off-line which deals with only a fixed sample or on-line which considers new information as it observed. Off-line change point problems deal with fixed sample sizes which are first observed and then detection and estimation of change points are done. [5] introduced the change point problem within the off-line setting. Since this pioneering work, methodologies used for change point detection have been widely researched on with methods extending to techniques for higher order moments within time series data. Ideally, it is desired to test how many change points are present within a given set of data and to estimate the parameters associated with each segment. If τ is known then the two samples only need to be compared. However, if τ is unknown then it has to be analyzed through change point analysis that entails both detection and estimation of the change point/change time. The null hypothesis of no change against the alternative that there exists a time when the distribution characteristics of the series changed is then tested. Stationarity in the strict sense, implies time-invariance of the distribution underlying the process.

The hypotheses would be stated as:

$$\begin{aligned} H_0 &: F\left(x;\theta\right) = F\left(x;\theta_1\right) \quad \text{for } t = 1,\cdots,n \\ H_1 &: F\left(x;\theta\right) = \begin{cases} F\left(x;\theta_1\right) & \text{for } t = 1,\cdots,\tau \\ F\left(x;\theta_2\right) & \text{for } t = \tau+1,\cdots,n \end{cases} \end{aligned} \tag{1}$$

The null hypothesis postulates that the distribution remains unchanged throughout within the sample of size n whereas the alternative postulates no change as in the null up to time τ when change occurs. Then the change point problem is to test the hypotheses about the population parameter(s)

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_n \text{ versus } H_1 : \theta_1 = \cdots = \theta_\tau \neq \theta_{\tau+1} = \cdots = \theta_n \tag{2}$$

where τ is unknown and needs to be estimated. If $\tau < n$ then the process distribution has changed and τ is referred to as the change point. We assume that there exists $\lambda \in [0,1]$ such that τ satisfies

$$\tau = \lambda n \tag{3}$$

where n is the number of observations in a given data set. Then hypothesis 2 can be restated as

$$\begin{aligned} H_0 &: \tau = n, \ (\lambda = 1) \\ H_1 &: \tau < n, \ (0 < \lambda < 1) \end{aligned} \tag{4}$$

At a given level of significance, if the null hypothesis is rejected, then the process X is said to be locally piecewise-stationary and can be approximated by a sequence of stationary processes that may share certain features such as the general functional form of the distribution F. Many authors such as [6]-[11] have considered both parametric and non-parametric methods of change point detection in time series data. Ideally, change points cannot be assumed to be known in advance hence the need for various methods of detection and estimation.

This paper is organized as follows: Section 2 gives an overview of the change point estimator based on a pseudo-distance measure. Section 3 provides key results for consistency of the estimator. Section 4 provides an application of the change point estimator to the shape and scale parameters of the generalized Pareto distribution. Section 5 gives an application of the estimator and consistency is shown through simulations. Finally 6 provides concluding remarks.

2. Change Point Estimator

The change point problem is addressed by using a "distance" function between distributions to describe the change. Given a distance function, a test statistic is constructed to guarantee a distance $> \epsilon (\epsilon \geq 0)$ between any two distributions based on a sample size n. Consider a given parametric model $f_\theta : \theta \in \Theta$ where Θ is the parameter space defined on a data set of size n. Let $X_1, \cdots, X_n$ be random variables and have probability densities $f(x;\theta_1), \cdots, f(x;\theta_n)$ with respect to σ-finite measure μ with $F(x;\theta)$ generating distinct measures if $\theta \in \Theta$

Definition 2.1 (ϕ-divergence). Let F_{θ_1} and F_{θ_2} be two probability distributions. Define the ϕ-divergence between the two distributions as

$$D_\phi (F_{\theta_1}, F_{\theta_2}) = D_\phi (\theta_1, \theta_2)$$

The broader family of ϕ-divergences that take the general form

$$D_\phi (\theta_1, \theta_2) = \int \phi \left(\frac{\mathrm{d}F_{\theta_1}}{\mathrm{d}F_{\theta_2}} \right) \mathrm{d}F_{\theta_2}$$

$$\begin{aligned} &= \int f_{\theta_2}(x)\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)\mathrm{d}\mu(x) \\ &= E_{\theta_2}\left[\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)\right],\ \phi\in\Phi \end{aligned} \tag{5}$$

where Φ is the class of all convex functions $\phi(t), t>0$ satisfying $\phi(1)=0, \phi''(1)>0$.

Assumption 1. The function $\phi\in\Phi:[0,\infty)\to(-\infty,+\infty)$ is convex and continuous. The restriction on $[0,\infty)$ is finite, twice continuously differentiable with $\phi(1)=\phi'(1)=0, \phi''(1)=1$.

At any point $t=0$, to avoid indeterminate expressions [12] gives the following assumptions in relation to the functions ϕ involved in the general definition of ϕ-divergence statistics,

$$\begin{aligned} &0\phi\left(\frac{0}{0}\right)=0 \\ &0\phi\left(\frac{p}{0}\right)=\lim_{u\to\infty}\frac{\phi(u)}{u} \end{aligned} \tag{6}$$

These assumptions ensure the existence of the integrals. Different choices of φ result in many divergences that play important roles in statistics including the Kullback-Leibler $\phi(t)=-\ln(t)$, total variation $\phi(t)=|t-1|$ among others. $D_\phi(\theta_1,\theta_2)\neq D_\phi(\theta_2,\theta_1)$ hence divergence measures are not distance measures but give some difference between two probability measures hence the term "pseudo-distance". More generally a divergence measure is a function of two probability density (or distribution) functions, which has non-negative values and takes the value zero only when the two arguments (distributions) are the same. A divergence measure grows larger as two distributions are further apart. Hence, a large divergence implies departure from the null hypothesis.

Generally, a change point problem's objective would be to propose an estimator for the possible change-point τ given a set of random variables.

Based on the divergence in 5 then a change point estimator can be constructed as;

$$D_{n\tau}=\max_{1<\tau<n}\left(\lambda(1-\lambda)\right)\frac{2}{\phi''(1)}D_\phi(\theta_1,\theta_2) \tag{7}$$

where $\lambda=\frac{\tau}{n}\in\Lambda:\Lambda=[0,1]$ and $\widehat{\theta}_1,\widehat{\theta}_2$ are the maximum likelihood estimates of the parameters before and after the change point.

To test for the possibility of having a change in distribution of $x_1,\cdots,x_n$ it is natural to compare the distribution function of the first τ observations to that of the last $(n-\tau)$ since the location of the change time is unknown. When τ is near the boundary points, say near 1 or near n then we are required to compare an estimation calculated on a correct large number of observations $(n-\tau)$ to an estimation from a small number of observations τ. This may result to an erratic

behavior of the test statistic [7] due to instability of the estimators of the parameters. If λ is not bounded away from zero and one, then the test statistic does not converge in distribution *i.e.* the critical values for the test statistic diverge to infinity as $n \to \infty$ to obtain a sequence of level α tests [13]. However, fixed critical values can be obtained for increasing sample sizes when λ is bounded away from zero and one and yields significant power gains if the change point is in Λ.

Let $\varepsilon > 0$ be small enough such that $\lambda \in (\epsilon, 1-\epsilon)$

Suppose that λ maximizes the test statistic over [0, 1] then under the null hypothesis,

$$\begin{aligned} &\sup_{\lambda\in(\epsilon,1-\epsilon)} D(\lambda) = O_p(1) \quad \forall\epsilon \\ &\sup_{\lambda\in[0,1]} D(\lambda) \to \infty \quad \text{as } n \to \infty \end{aligned} \tag{8}$$

[13]. By this result and for $N(\epsilon) = \epsilon n, \cdots, (1-\epsilon)n$ then the test statistic becomes,

$$D_{n\tau} = \max_{\tau\in N(\epsilon)} \left(\frac{\tau}{n}\left(1-\frac{\tau}{n}\right)\right) \frac{2}{\phi''(1)} D_\phi(\theta_1, \theta_2) \tag{9}$$

The change-point estimator $\hat{\tau}$ of a change point τ is the point at which there is maximal sample evidence for a change in distributional parameters characterized by maximum divergence. It is estimated by the least value of τ that maximizes the test statistic 9.

$$\hat{\tau} = \min\left\{ \tau : D_{n\tau} = \max_{\tau\in N(\epsilon)} \left(\frac{\tau}{n}\left(1-\frac{\tau}{n}\right)\right) \frac{2}{\phi''(1)} D_\phi(\theta_1, \theta_2) \right\} \tag{10}$$

3. Consistency of the Change Point Estimator

A minimal requirement for a good statistical decision rule is its increasing reliability with increasing sample sizes [14].

Let $x_1, \cdots, x_n$ be a sample of fixed size n with the density function $f(x;\theta)$ for $\theta \in \Theta \subset R^d$ and $L(x;\theta)$ be the likelihood function. It can be shown that by Taylor's theorem under the null hypothesis, the ϕ, divergence based estimator can be reduced to a two-sample Wald-type test statistic of the form

$$\widehat{W_{n\tau}} = \max_{\tau\in N(\epsilon)} \left(\frac{\tau}{n}\left(1-\frac{\tau}{n}\right)\right) \left(\hat{\theta}_1 - \theta_1\right) I(\theta_0) \left(\hat{\theta}_2 - \theta_2\right) \tag{11}$$

Suppose $x_1, \cdots, x_n$ are iid random variables of size n with probability density function $f(x;\theta)$ with $\theta = (\theta_1, \cdots, \theta_k)', k < n$ being the vector of parameters governing the pdf. The likelihood function can be expressed as

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i;\theta) \tag{12}$$

It is more convenient to work with the logarithm of the likelihood function given by

$$\ell(\theta \mid x) = \sum_{i=1}^{n} \log f(x_i;\theta) \tag{13}$$

Since the logarithm is a monotone increasing function, maximizing the likelihood function is equivalent to maximizing the log-likelihood function. Introduce the following notations:

$$\nabla_\theta \log f\left(x_i;\theta\right)=\frac{\partial}{\partial\theta}\log f\left(x_i;\theta\right) \tag{14}$$

$$\nabla_\theta^2 \log f\left(x_i;\theta\right)=\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f\left(x_i;\theta\right) \tag{15}$$

$$H_n\left(\theta\right)=\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2 \log f\left(x_i;\theta\right) \tag{16}$$

$$U_{jm}\left(\theta\right)=\sum_{i=j}^{m}\nabla_\theta \log f\left(x_i;\theta\right), 1\le j\le m\le n \tag{17}$$

The following equalities hold as $n\to\infty$.

$$\begin{aligned} &H_n\left(\theta\right)\to -I\left(\theta\right)\\ &H_n\left(\theta\right)+\frac{1}{n}I\left(\theta\right)\to 0\end{aligned} \tag{18}$$

On assumption that $\theta_1\neq\theta_2$ for $\theta_1,\theta_2\in\Theta\subset R^d$, then

$$\begin{aligned} &\widehat{\theta}_1\to\theta_1,\widehat{\theta}_2\to\theta_2 \text{ as } n\to\infty\\ &\widehat{\theta}_1 \text{ and } \widehat{\theta}_2 \text{ are solutions to}\\ &\sum_{i=1}^{\tau}\nabla_\theta \log f\left(x_i;\theta\right)=0 \text{ and } \sum_{i=\tau+1}^{n}\nabla_\theta \log f\left(x_i;\theta\right)=0 \text{ respectively.}\end{aligned} \tag{19}$$

Theorem 3.1. Let $0<\delta_1<\delta_2<\infty, n_1=n\delta_1, n_2=n\delta_2$

$$\lim_{\tau\to\infty}\max\left\{\left\|\tau^{-\frac{1}{2}}U_\tau\left(\theta_0\right)\right\|: n_1<\tau<n_2\right\}=O_p\left(1\right) \tag{20}$$

Theorem 3.2. Let $0<\epsilon<1-\epsilon$ for $\epsilon>0$ small enough. Then as $n\to\infty$

$$\max\left\{\left\|\tau^{-\frac{1}{2}}I\left(\theta_0\right)^{-1}U_{n\tau}\left(\theta_0\right)\right\|: \tau\in N\left(\epsilon\right)\right\}=O_p\left(1\right) \tag{21}$$

For the proof of theorems 3.1 and 3.2 see [15].

Theorem 3.3. Let $0<\delta_1<\delta_2<\infty$ and $n_1=n\delta_1, n_2=n\delta_2$ For $n\ge 1$

$$\lim_{n\to\infty}\left\{\max\left[n^{1/2}\left\|\left(\widehat{\theta}_n-\theta_0\right)-\frac{1}{n}I\left(\theta_0\right)^{-1}U_n\left(\theta_0\right)\right\|: n_1<n_2\right]\right\}\to 0 \tag{22}$$

Proof

$$\begin{aligned} 0=U_n\left(\widehat{\theta}_n\right)&=U_n\left(\theta_0\right)+\left(\widehat{\theta}_n-\theta_0\right)U_n'\left(\theta_0\right)+\frac{1}{2}\left(\widehat{\theta}_n-\theta_0\right)^2U_n'''\left(\theta_0\right)\\ &=U_n\left(\theta_0\right)-nH_n\left(\theta_0\right)+\tilde{R}\\ &=n^{-\frac{1}{2}}U_n\left(\theta_0\right)-nH_n\left(\theta_0\right)n^{-\frac{1}{2}}\left(\widehat{\theta}_n-\theta_0\right)+n^{-\frac{1}{2}}\tilde{R}\end{aligned} \tag{23}$$

The third term on the RHS is $o_p\left(1\right)$ [14]. By definition of MLE $U_n\left(\widehat{\theta}_n\right)=0$.

$$U_n\left(\theta_0\right)=-nH_n\left(\theta_n\right)\left(\widehat{\theta}_n-\theta_0\right)n^{-\frac{1}{2}}U_n\left(\theta_0\right)=-n^{\frac{1}{2}}H_n\left(\theta_0\right)\left(\widehat{\theta}_n-\theta_0\right) \tag{24}$$

From Equation (24) we obtain

$$n^{\frac{1}{2}}\left(\widehat{\theta}_n-\theta_0\right)=-n^{\frac{1}{2}}H_n\left(\theta_0\right)^{-1}U_n\left(\theta_0\right)$$

Hence

$$n^{\frac{1}{2}}\left\|\left(\widehat{\theta}_n-\theta_0\right)-\frac{1}{n}I\left(\theta_0\right)^{-1}U_n\left(\theta_0\right)\right\|\le\left\|-H_n\left(\theta_0\right)^{-1}I\left(\theta_0\right)\right\|\left\|n^{-\frac{1}{2}}U_n\left(\theta_0\right)\right\| \quad (25)$$

But by Equation (18)

$$\left\|-H_n\left(\theta_0\right)^{-1}I\left(\theta_0\right)\right\|\to 0$$

By theorem 3.1 $\left\|n^{-\frac{1}{2}}U_n\left(\theta_0\right)\right\|$ is bounded in probability. Hence the proof.

Theorem 3.4. Let $0<\epsilon<1-\epsilon$ for $\epsilon>0$ small enough. Then

$$\lim_{n\to\infty}\left\{\max\left[\tau^{1/2}\left\|\left(\widehat{\theta}_1-\widehat{\theta}_2\right)-\frac{n}{\tau\left(n-\tau\right)}I\left(\theta_0\right)^{-1}U_{n\tau}\left(\theta_0\right)\right\|:\tau\in N\left(\epsilon\right)\right]\right\}\to 0 \quad (26)$$

Proof

$$\begin{aligned}&\left(\widehat{\theta}_1-\widehat{\theta}_2\right)-\frac{n}{\tau\left(n-\tau\right)}I\left(\theta_0\right)^{-1}U_{n\tau}\left(\theta_0\right)\\&=\left(\widehat{\theta}_1-\theta_0-\frac{1}{\tau}I\left(\theta_0\right)^{-1}U_{1\tau}\left(\theta_0\right)\right)-\left(\widehat{\theta}_2-\theta_0-\frac{1}{n-\tau}I\left(\theta_0\right)^{-1}U_{\tau+1,n}\left(\theta_0\right)\right)\\&\max\tau^{\frac{1}{2}}\left(\widehat{\theta}_1-\widehat{\theta}_2\right)-\frac{n}{\tau\left(n-\tau\right)}I\left(\theta_0\right)^{-1}U_{n\tau}\left(\theta_0\right) \quad (27)\\&\le\max\tau^{\frac{1}{2}}\left(\widehat{\theta}_1-\theta_0-\frac{1}{\tau}I\left(\theta_0\right)^{-1}U_{1\tau}\left(\theta_0\right)\right)\\&\quad-\max\tau^{\frac{1}{2}}\left(\widehat{\theta}_2-\theta_0-\frac{1}{n-\tau}I\left(\theta_0\right)^{-1}U_{\tau+1,n}\left(\theta_0\right)\right)\end{aligned}$$

Considering the term on the RHS, by theorem 3.3. For $\tau,n\to\infty$

$$\max\tau^{\frac{1}{2}}\left(\widehat{\theta}_1-\theta_0-\frac{1}{\tau}I\left(\theta_0\right)^{-1}U_{1\tau}\left(\theta_0\right)\right)\to 0$$

$$\max\tau^{\frac{1}{2}}\left(\widehat{\theta}_2-\theta_0-\frac{1}{n-\tau}I\left(\theta_0\right)^{-1}U_{\tau+1,n}\left(\theta_0\right)\right)\to 0$$

Hence the proof.

Assume that within a finite set of data a change point τ exists and $n\to\infty$ such that $\tau,\left(n-\tau\right)\to\infty$

Define,

$$U_{n\tau}\left(\theta_0\right)=U_{1\tau}\left(\theta_0\right)-\frac{\tau}{n}U_{1n}\left(\theta_0\right)$$

Consider the following two sample homogeneity test

$$Q_{n\tau}=\frac{n}{\tau\left(n-\tau\right)}U_{n\tau}\left(\theta_0\right)I\left(\theta_0\right)U_{n\tau}\left(\theta_0\right)\ \text{ for } 1<\tau<n \quad (28)$$

[15] defined a consistent estimate of 28 as

$$\widehat{Q_{n\tau}} = \frac{n}{\tau(n-\tau)} U_\tau\left(\widehat{\theta_n}\right)\left(H_n\left(\widehat{\theta_n}\right)\right)^{-1} U_\tau\left(\widehat{\theta_n}\right) \tag{29}$$

By the principles of maximum likelihood estimation, $Q_{nn} = 0$ since $U_n\left(\widehat{\theta_n}\right) = 0$, $Q_{n0} = 0$ since $U_0(.) = 0$.

Consider $U_\tau\left(\widehat{\theta_n}\right)$. By Taylor's theorem,

$$\begin{aligned} U_\tau\left(\widehat{\theta_n}\right) &= U_\tau(\theta_\tau) + \left(\widehat{\theta_n} - \widehat{\theta_\tau}\right) U'_\tau(\theta_\tau) \\ U'_\tau(\theta_\tau) &= \sum_{i=1}^{\tau} \frac{\partial^2}{\partial \theta_\tau^2} \log f(x;\theta_\tau) \\ &= \left(\widehat{\theta_n} - \widehat{\theta_\tau}\right) \sum_{i=1}^{\tau} \frac{\partial^2}{\partial \theta_\tau^2} \log f(x;\theta_\tau) \\ &= \tau H_\tau\left(\widehat{\theta_\tau}\right)\left(\widehat{\theta_n} - \widehat{\theta_\tau}\right) \end{aligned} \tag{30}$$

Since by the principle of maximum likelihood estimation $U_\tau(\theta_\tau) = 0$.

$$U_\tau\left(\widehat{\theta_n}\right) = -\tau H_\tau\left(\widehat{\theta_\tau}\right)\left(\widehat{\theta_\tau} - \widehat{\theta_n}\right) \tag{31}$$

$$\begin{aligned} \left(\widehat{\theta_\tau} - \widehat{\theta_n}\right) &= -\tau\left\{H_\tau\left(\widehat{\theta_\tau}\right)\right\}^{-1} U_\tau\left(\widehat{\theta_n}\right) \\ &= -\tau\left\{H_\tau\left(\widehat{\theta_\tau}\right)\right\}^{-1}\left\{U_{1\tau}\left(\widehat{\theta_0}\right) - \frac{\tau}{n} U_{1n}\left(\widehat{\theta_0}\right)\right\} \\ &= -\tau\left\{H_\tau\left(\widehat{\theta_\tau}\right)\right\}^{-1}\left\{\frac{1}{\tau} U_{1n}\left(\widehat{\theta_0}\right) - \frac{1}{n-\tau} U_{1n}\left(\widehat{\theta_0}\right)\right\} \\ &= \tau\left\{H_\tau\left(\widehat{\theta_\tau}\right)\right\}^{-1}\left\{\frac{1}{n-\tau} U_{1n}\left(\widehat{\theta_0}\right) - \frac{1}{\tau} U_{1n}\left(\widehat{\theta_0}\right)\right\} \\ &= -\frac{n\tau}{n-\tau}\left\{H_\tau\left(\widehat{\theta_\tau}\right)\right\}^{-1} U_{1n}\left(\widehat{\theta_0}\right) \\ &= \frac{n\tau}{n-\tau} I\left(\hat{\theta}_\tau\right) U_{1n}\left(\widehat{\theta_0}\right) \end{aligned} \tag{32}$$

By the CLT,

$$\left(\widehat{\theta_\tau} - \widehat{\theta_n}\right) \to N\left(0, \frac{n-\tau}{n\tau} I(\theta_\tau)^{-1}\right)$$

and thus $\left(\widehat{\theta_\tau} - \widehat{\theta_n}\right)$ has squared Mahalanobis norm

$$\left(\widehat{\theta_\tau} - \widehat{\theta_n}\right)\left(\frac{n-\tau}{n\tau} I(\theta_\tau)^{-1}\right)^{-1}\left(\widehat{\theta_\tau} - \widehat{\theta_n}\right) \tag{33}$$

Hence

$$\left(\widehat{\theta_\tau} - \widehat{\theta_n}\right)\left(\frac{n-\tau}{n\tau} I(\theta_\tau)^{-1}\right)^{-1}\left(\widehat{\theta_\tau} - \widehat{\theta_n}\right) \approx \widehat{Q_{n\tau}} \tag{34}$$

implying that $Q_{n\tau}$ is approximately equal to the Mahalanobis norm of $\left(\widehat{\theta_\tau} - \widehat{\theta_n}\right)$. The Mahalanobis norm can be used to detect change points within a given finite time series data [11]. Since the test statistic 33 can quantify the difference between $\left(\widehat{\theta_n}\right)$ and $\left(\widehat{\theta_\tau}\right)$ then $\widehat{Q_{n\tau}}$ can similarly be used to quantify the deviation

between the two parameter estimates. The value of $\widehat{Q_{n\tau}}$ ideally grows larger in evidence of the alternative hypothesis and tends towards zero when the null hypothesis is true. Suppose we define a maximal type test statistic $\widehat{Q_{n\tau}}(t)$ such that

$$\widehat{Q_{n\tau}} = \max\left\{\widehat{Q_{n\tau}}(t) : t \in N(\epsilon)\right\} \tag{35}$$

then we can obtain a measure of the largest difference between $\left(\widehat{\theta_n}\right)$ and $\left(\widehat{\theta_\tau}\right)$. Consider the divergence based estimator which was reduced to a two sample test statistic in Equation (11).

Definition 3.1. A matrix M is called positive definite if $x'Mx \geq 0, \forall x \in R^n$, with equality if and only if $x = 0$. The following inequality holds,

$$\left\|x'Mx - y'My\right\| \leq |M|\left\{\|x-y\|^2 + 2\|y\|\|x-y\|\right\} \tag{36}$$

Consider the following result

$$\begin{aligned}\left|W_{n\tau} - Q_{n\tau}\right| = &\left|\frac{\tau(n-\tau)}{n}\left(\widehat{\theta_1} - \widehat{\theta_2}\right) I(\theta_0)\left(\widehat{\theta_1} - \widehat{\theta_2}\right)\right. \\ &\left. - \left\{\frac{n}{\tau(n-\tau)} I(\theta_0)^{-1} U_{n\tau}(\theta_0)\right\}' I(\theta_0)\left\{\frac{n}{\tau(n-\tau)} I(\theta_0)^{-1} U_{n\tau}(\theta_0)\right\}\right|\end{aligned} \tag{37}$$

By inequality 36

$$\begin{aligned}\left|W_{n\tau} - Q_{n\tau}\right| < &\left|I(\theta_0)\right|\left\{\left\|\left(\widehat{\theta_1} - \widehat{\theta_2}\right) - \frac{n}{\tau(n-\tau)} I(\theta_0)^{-1} U_{n\tau}(\theta_0)\right\|^2\right. \\ &\left. + 2\left\|\frac{n}{\tau(n-\tau)} I(\theta_0)^{-1} U_{n\tau}(\theta_0)\right\|\left\|\left(\widehat{\theta_1} - \widehat{\theta_2}\right) - \frac{n}{\tau(n-\tau)} I(\theta_0)^{-1} U_{n\tau}(\theta_0)\right\|\right\}\end{aligned} \tag{38}$$

Consider the last term on the RHS. By the result of theorem 3.4

$$\left\|\left(\widehat{\theta_1} - \widehat{\theta_2}\right) - \frac{n}{\tau(n-\tau)} I(\theta_0)^{-1} U_{n\tau}(\theta_0)\right\| \to 0 \tag{39}$$

And hence

$$\left\|\left(\widehat{\theta_1} - \widehat{\theta_2}\right) - \frac{n}{\tau(n-\tau)} I(\theta_0)^{-1} U_{n\tau}(\theta_0)\right\|^2 \to 0 \tag{40}$$

Considering the second term on the RHS. By the results in theorem 3.2,

$$\left\|I(\theta_0)^{-1} U_{n\tau}(\theta_0)\right\| = O_p(1) \tag{41}$$

From these results as $\tau, n \to \infty$

$$\left|W_{n\tau} - Q_{n\tau}\right| \to 0 \tag{42}$$

Definition 3.2. (Asymptotic consistency). A change point detection algorithm is said to be asymptotically consistent if the estimated segmentation is such that

$$\max\left|\frac{\hat{\tau}}{n} - \frac{\tau}{n}\right| \to 0 \tag{43}$$

The change point fractions are consistent, and not the indexes themselves. Consistency results in the literature only deal with change point fractions since the distances $\left|\hat{\tau}-\tau\right|$ and their estimated counter parts do not converge to zero [11].

4. Change Point Analysis in the Generalized Pareto Distribution

Definition 4.1. The Generalized Pareto distribution function is defined by;

$$H\left(x\right)=\begin{cases}1-\left(1+\dfrac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}}, & \xi\neq 0\\ 1-\exp\left(-\dfrac{x}{\sigma}\right), & \xi=0\end{cases} \tag{44}$$

where,

$$x\in\begin{cases}\left[0,\infty\right), & \xi\geq 0\\ \left[0,-\dfrac{\sigma}{\xi}\right], & \xi<0\end{cases}$$

σ is referred to as the scale parameter characterizes the spread of the distribution and ξ referred to as the tail index/shape parameter determines the tail thickness. More specifically, given that $X\sim GP\left(\sigma,\xi\right)$ then the probability density function is;

$$h\left(x\right)=\begin{cases}\dfrac{1}{\sigma}\left(1+\dfrac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}-1}, & \xi\neq 0\\ \dfrac{1}{\sigma}\exp\left(-\dfrac{x}{\sigma}\right), & \xi=0\end{cases} \tag{45}$$

For any given finite set of data, at least one of the following is likely at any given change point $\tau\left(1<\tau<n\right)$: ξ changes by a non-zero quantity; σ changes by a non-zero quantity; both ξ and σ change by non-zero quantities. A simple change point problem can be formulated in one of the following ways;

$$\begin{aligned}&H_0: X_t\sim GP\left(\sigma_1,\xi_1\right) \text{ against}\\ &H_1: X_t\sim GP\left(\sigma_1,\xi_1\right) \quad t\leq\tau\\ &\qquad X_t\sim GP\left(\sigma_2,\xi_2\right) \quad t>\tau\end{aligned} \tag{46}$$

$$\begin{aligned}&H_0: X_t\sim GP\left(\sigma_1,\xi_1\right) \text{ against}\\ &H_1: X_t\sim GP\left(\sigma_1,\xi_1\right) \quad t\leq\tau\\ &\qquad X_t\sim GP\left(\sigma_1,\xi_2\right) \quad t>\tau\end{aligned} \tag{47}$$

$$\begin{aligned}&H_0: X_t\sim GP\left(\sigma_1,\xi_1\right) \text{ against}\\ &H_1: X_t\sim GP\left(\sigma_1,\xi_1\right) \quad t\leq\tau\\ &\qquad X_t\sim GP\left(\sigma_2,\xi_1\right) \quad t>\tau\end{aligned} \tag{48}$$

Since change points are unknown in advance, then either of the three hypothesis formulations is likely. Without knowledge on the types of changes con-

tained in the time series, the question arises on which testing procedure to use. In most instances hypotheses 46 is tested since it is assumed that both distributional parameters change.

Figure 1 shows different GP density plots with a constant scale parameter but varying shape parameters. On the other hand, Figure 2 shows different GP density plots with a both scale and shape parameters varying. If any of the parameters were to change at any given point in time, then the thickness of the general tail distribution would change and this would in turn have an effect of the intensity of extreme values observed.

Assume that X is independently and identically distributed random variables drawn for the generalized Pareto distribution and consider a sample data set $x_1, \cdots, x_\tau, x_{\tau+1}, \cdots, x_n$ of fixed size $n (n \geq 3)$. Say f_{θ_1} is governed by the parameter space $\theta_1 = (\xi_1, \sigma_1)$ and f_{θ_2} is governed by the parameter space $\theta_2 = (\xi_2, \sigma_2)$ where $\theta_1 \neq \theta_2 \in \Theta$. The data set is assumed to contain an unknown change point τ where the distribution parameters ξ and σ abruptly change. Then

$$x_1, \cdots, x_\tau \sim f_{\theta_1}(x)$$
$$x_{\tau+1}, \cdots, x_n \sim f_{\theta_2}(x)$$

Then the density function 49 governs the first τ observations and 50 governs the last $(n-\tau)$ observations.

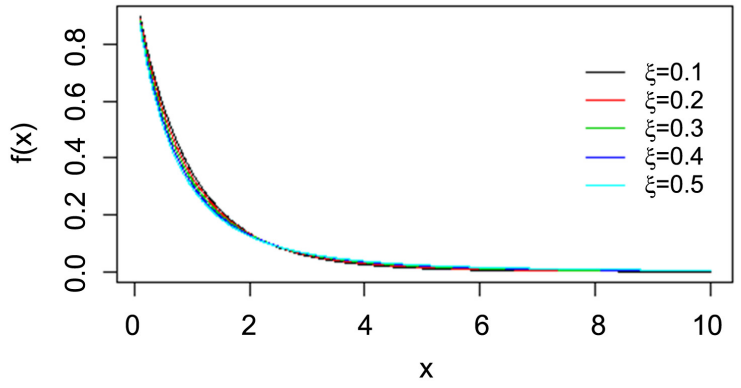

Figure 1. Density plot with constant scale.

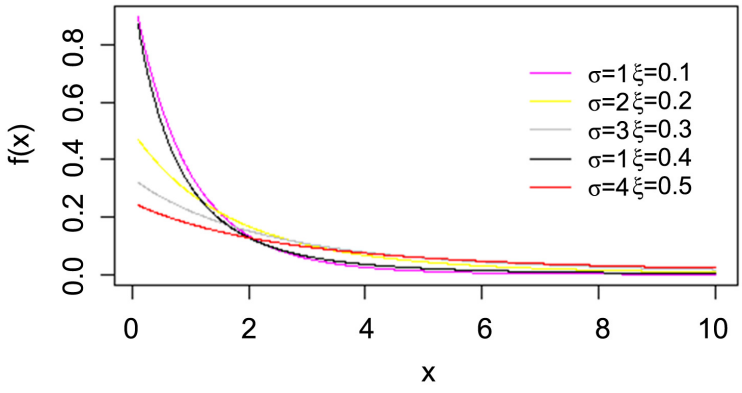

Figure 2. Density plot with varying scale and shape.

$$f_{\theta_1}(x)=\begin{cases}\dfrac{1}{\sigma_1}\left(1+\dfrac{\xi_1 x}{\sigma_1}\right)^{-\frac{1}{\xi_1}-1}, & \xi_1\neq 0\\ \dfrac{1}{\sigma_1}\exp\left(-\dfrac{x}{\sigma_1}\right), & \xi_1=0\end{cases} \tag{49}$$

$$f_{\theta_2}(x)=\begin{cases}\dfrac{1}{\sigma_2}\left(1+\dfrac{\xi_2 x}{\sigma_2}\right)^{-\frac{1}{\xi_2}-1}, & \xi_1\neq 0\\ \dfrac{1}{\sigma_2}\exp\left(-\dfrac{x}{\sigma_2}\right), & \xi_1=0\end{cases} \tag{50}$$

We will restrict to the case where $\xi>0$ *i.e.* heavy tailed distributions thereby only considering the first part of the density function with support $x\in[0,\infty)$.

From the divergence in Equation (5), let $\phi(t)=-\log(t)$

$$\begin{aligned}D_{\phi}(\theta_1,\theta_2)&=\int f_{\theta_2}(x)\phi\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)\mathrm{d}\mu(x)\\&=\int f_{\theta_2}(x)-\log\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)\mathrm{d}\mu(x)\\&=\int f_{\theta_2}(x)\log\left(\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}\right)\mathrm{d}\mu(x)\\&=D_{KL}\left(f_{\theta_2}(x),f_{\theta_1}(x)\right)\equiv D_{KL}(\theta_2,\theta_1)\end{aligned} \tag{51}$$

An application of properties of the generalized Pareto distribution [16], numerical computations and methods of integration the divergence between two generalized Pareto distributions becomes

$$\begin{aligned}D_{KL}(\theta_2,\theta_1)&=\log\left(\frac{\sigma_2}{\sigma_1}\right)-(1+\xi_1)\\&\quad-\left(\frac{1}{\xi_2}+1\right)\frac{\sigma_1}{\sigma_2}\xi_1\int\left(1+\frac{\xi_1}{\sigma_1}x\right)^{-\frac{1}{\xi_1}}\left(\frac{\xi_2}{\sigma_2}+1\right)^{-1}\mathrm{d}x\end{aligned} \tag{52}$$

The divergence is a function of the parameters of the two densities.

5. Simulation Study

The performance of the estimator is examined by considering the effects of the change in sample size. The single change-point estimation problem is considered where the change-point τ is fixed at $n/2$ for n = 200, 500, 1000. **Figures 3-5** display the plots for the location of the change-point estimator as estimated by the proposed estimator 10 with the divergence measure as in 52 for the various sample sizes. The hypothesis considered here is

$$\begin{aligned}&H_0: X_t\sim GP(1,0.1)\quad\text{against}\\&H_1: X_t\sim GP(1,0.1)\quad t\leq\tau\\&\qquad X_t\sim GP(3,0.35)\quad t>\tau\end{aligned} \tag{53}$$

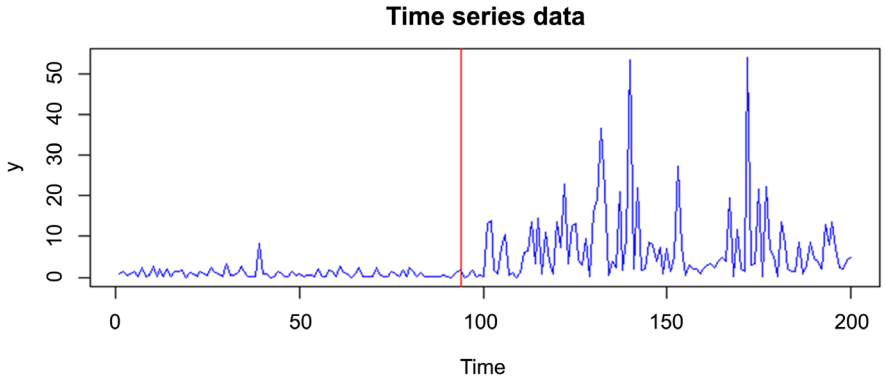

Figure 3. Sample size= 200, $\tau = 100$, $\hat{\tau} = 88$.

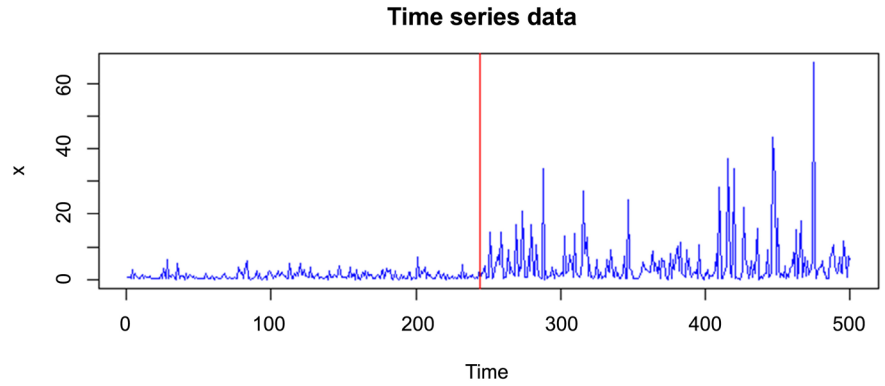

Figure 4. Sample size= 500, $\tau = 250$, $\hat{\tau} = 245$.

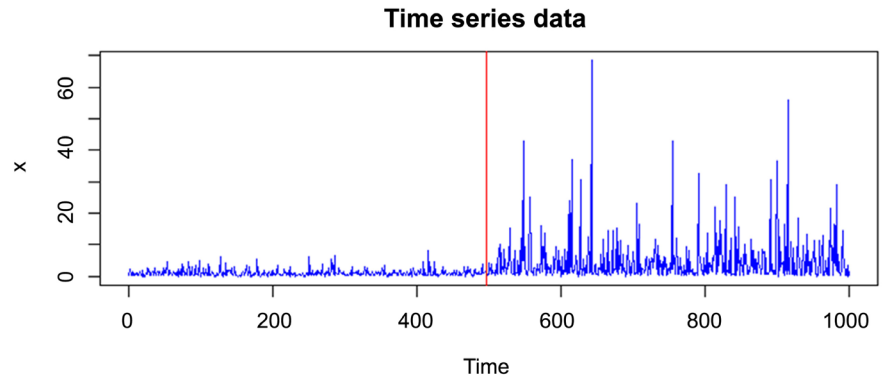

Figure 5. Sample size= 1000, $\tau = 500$, $\hat{\tau} = 494$.

To check consistency of the estimator, we consider the following: first, we consider data simulated from the GP density with parameters $(1, \xi_1)$ and $(3, \xi_2)$ for the scale and shape respectively before and after the change point. 1000 simulations are carried out to estimate the change point and the results are given in Table 1 and Table 2.

6. Conclusion

In this paper, a divergence (pseudo-distance) based estimator is used to detect change points within a parametric framework focusing on the generalized Pareto

Table 1. Effect of the sample size with varying scale and varying shape ($\tau = n/2$).

		$\xi_2 = 0.4$		$\xi_2 = 0.3$		$\xi_2 = 0.2$	
n	τ	$\hat{\tau}$	$\left\lvert\frac{\hat{\tau}}{n} - \frac{\tau}{n}\right\rvert$	$\hat{\tau}$	$\left\lvert\frac{\hat{\tau}}{n} - \frac{\tau}{n}\right\rvert$	$\hat{\tau}$	$\left\lvert\frac{\hat{\tau}}{n} - \frac{\tau}{n}\right\rvert$
100	50	35	0.15	31	0.19	30	0.2
200	100	88	0.06	85	0.075	82	0.09
500	250	245	0.01	241	0.018	241	0.018
1000	500	494	0.006	493	0.007	491	0.009

Table 2. Effect of the sample size with varying scale and varying shape ($\tau = n/3$).

		$\xi_2 = 0.4$		$\xi_2 = 0.3$		$\xi_2 = 0.2$	
n	τ	$\hat{\tau}$	$\left\lvert\frac{\hat{\tau}}{n} - \frac{\tau}{n}\right\rvert$	$\hat{\tau}$	$\left\lvert\frac{\hat{\tau}}{n} - \frac{\tau}{n}\right\rvert$	$\hat{\tau}$	$\left\lvert\frac{\hat{\tau}}{n} - \frac{\tau}{n}\right\rvert$
200	66	70	0.02	69	0.015	64	0.01
500	166	164	0.004	163	0.006	165	0.002
1000	333	330	0.003	333	0	333	0

distribution. Change points are attributed to the change in model parameters at unknown points in time with the parameter estimates before and after the change point unknown. The estimator is shown to be consistent theoretically. Simulation studies also show that the change point estimator is consistent.

Acknowledgements

The first author thanks the Pan-African University Institute of Basic Sciences, Technology and Innovation (PAUSTI) for funding this research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Brodsky, E. and Darkhovsky, B.S. (2013) Nonparametric Methods in Change Point Problems. Vol. 243, Springer Science & Business Media, Berlin.

[2] Killick, R. and Eckley, I. (2014) Change Point: An R Package for Change Point Analysis. *Journal of Statistical Software*, **58**, 1-19. https://doi.org/10.18637/jss.v058.i03

[3] Korkas, K.K. and Fryzlewicz, P. (2017) Multiple Change-Point Detection for Non-Stationary Time Series Using Wild Binary Segmentation. *Statistica Sinica*, **27**, 287-311. https://doi.org/10.5705/ss.202015.0262

[4] Csörgö, M. and Horváth, L. (1997) Limit Theorems in Change-Point Analysis. Vol. 18, John Wiley & Sons Inc., Hoboken.

[5] Page, E. (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

[6] Cheng, L., AghaKouchak, A., Gilleland, E. and Katz, R.W. (2014) Non-Stationary Extreme Value Analysis in a Changing Climate. *Climatic Change*, **127**, 353-369. https://doi.org/10.1007/s10584-014-1254-5

[7] Jarušková, D. and Rencová, M. (2008) Analysis of Annual Maximal and Minimal Temperatures for Some European Cities by Change Point Methods. *Environmetrics*, **19**, 221-233. https://doi.org/10.1002/env.865

[8] Naveau, P., Guillou, A. and Rietsch, T. (2014) A Non-Parametric Entropy-Based Approach to Detect Changes in Climate Extremes. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **76**, 861-884. https://doi.org/10.1111/rssb.12058

[9] Dupuis, D., Sun, Y. and Wang, H.J. (2015) Detecting Change-Points in Extremes. *Statistics and Its Interface*, **8**, 19-31. https://doi.org/10.4310/SII.2015.v8.n1.a3

[10] Dette, H. and Wu, W. (2018) Change Point Analysis in Non-Stationary Processes: A Mass Excess Approach.

[11] Truong, C., Oudre, L. and Vayatis, N. (2018) A Review of Change Point Detection Methods.

[12] Pardo, L. (2018) Statistical Inference Based on Divergence Measures. Chapman and Hall/CRC, London. https://doi.org/10.1201/9781420034813

[13] Andrews, D.W. (1993) Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica*: *Journal of the Econometric Society*, **61**, 821-856. https://doi.org/10.2307/2951764

[14] Sen, P.K. and Singer, J.M. (2017) Large Sample Methods in Statistics (1994): An Introduction with Applications. CRC Press, Boca Raton. https://doi.org/10.1201/9780203711606

[15] Hawkins Jr., D.L. (1983) Sequential Detection Procedures for Autoregressive Processes. Dept. of Statistics, Tech. Rep., North Carolina State University, Raleigh.

[16] Embrechts, P., Klüppelberg, C. and Mikosch, T. (2013) Modelling Extremal Events: For Insurance and Finance. Vol. 33, Springer Science & Business Media, Berlin.

Appendix

Derivation of the change point estimator W_{nr}

Consider a second order Taylor expansion of $D_\phi\left(\widehat{\theta}_1,\widehat{\theta}_2\right)$ about the true parameter values θ_1,θ_2

For $i=1,\cdots,d$

$$\begin{aligned}D_\phi\left(\widehat{\theta}_1,\widehat{\theta}_2\right)&=D_\phi\left(\theta_1,\theta_2\right)+\sum_{i=1}^{d}\frac{\partial D_\phi\left(\theta_1,\theta_2\right)}{\partial\theta_{1i}}\left(\widehat{\theta_{1i}}-\theta_{1i}\right)\\&+\sum_{i=1}^{d}\frac{\partial D_\phi\left(\theta_1,\theta_2\right)}{\partial\theta_{1i}}\left(\widehat{\theta_{1i}}-\theta_{1i}\right)\\&+\frac{1}{2}\sum_{i=1}^{d}\sum_{i=1}^{d}\frac{\partial^2 D_\phi\left(\theta_1,\theta_2\right)}{\partial\theta_{1i}\partial\theta_{1j}}\left(\widehat{\theta_{1i}}-\theta_{1i}\right)'\left(\widehat{\theta_{1i}}-\theta_{1i}\right)\\&+\frac{1}{2}\sum_{i=1}^{d}\sum_{i=1}^{d}\frac{\partial^2 D_\phi\left(\theta_1,\theta_2\right)}{\partial\theta_{2i}\partial\theta_{2j}}\left(\widehat{\theta_{2i}}-\theta_{2i}\right)'\left(\widehat{\theta_{2j}}-\theta_{2j}\right)\\&+\sum_{i=1}^{d}\sum_{i=1}^{d}\frac{\partial^2 D_\phi\left(\theta_1,\theta_2\right)}{\partial\theta_{1i}\partial\theta_{2j}}\left(\widehat{\theta_{1i}}-\theta_{1i}\right)'\left(\widehat{\theta_{2j}}-\theta_{2j}\right)\\&+o\left(\left\|\widehat{\theta_1}-\theta_1\right\|^2\right)+o\left(\left\|\widehat{\theta_2}-\theta_2\right\|^2\right)\end{aligned}\tag{54}$$

Under the assumption of the null hypothesis,

$$\begin{aligned}\frac{\partial D_\phi\left(\theta_1,\theta_2\right)}{\partial\theta_{1i}}&=\int\phi'\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1i}}\mathrm{d}\mu(x)\\&=\phi'(1)\int\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1i}}\mathrm{d}\mu(x)\\&=\phi'(1)\frac{\partial}{\partial\theta_{1i}}\int f_{\theta_1}(x)\mathrm{d}\mu(x)\\&=0\end{aligned}\tag{55}$$

This is by assumption 1 and that interchanges of derivatives and integrals are valid.

$$\begin{aligned}\frac{\partial^2 D_\phi\left(\theta_1,\theta_2\right)}{\partial\theta_{1i}\partial\theta_{1j}}&=\int\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1i}}\phi''\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1j}}\frac{1}{f_{\theta_2}(x)}\mathrm{d}\mu(x)\\&=\phi''(1)\int\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1i}}\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1j}}\frac{1}{f_{\theta_1}(x)}\mathrm{d}\mu(x)\\&=\int\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1i}}\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1j}}\frac{1}{f_{\theta_1}(x)}\mathrm{d}\mu(x)\\\frac{\partial^2 D_\phi\left(\theta_1,\theta_2\right)}{\partial\theta_{2i}\partial\theta_{2j}}&=\int\phi''\left(\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}\right)\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2i}}\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2j}}\frac{1}{f_{\theta_1}(x)}\mathrm{d}\mu(x)\\&=\phi''(1)\int\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2i}}\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2j}}\frac{1}{f_{\theta_1}(x)}\mathrm{d}\mu(x)\\&=\phi''(1)\int\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2i}}\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2j}}\frac{1}{f_{\theta_1}(x)}\mathrm{d}\mu(x)\end{aligned}\tag{56}$$

$$\begin{aligned}\frac{\partial^2 D_\phi(\theta_1,\theta_2)}{\partial\theta_{1i}\partial\theta_{2j}} &= \int \phi''\left(\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\right)\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1i}}\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2j}} - \frac{1}{\left(f_{\theta_2}(x)\right)^2} f_{\theta_1}(x)\mathrm{d}\mu(x)\\ &= -\phi''(1)\int \frac{\partial f_{\theta_1}(x)}{\partial\theta_{1i}}\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2j}}\frac{1}{f_{\theta_1}(x)}\mathrm{d}\mu(x)\\ &= -\left\{\frac{\partial^2 D_\phi(\theta_1,\theta_2)}{\partial\theta_{1i}\partial\theta_{1j}}\right\}\end{aligned}$$

By the standard regularity assumptions (theorem 5.2.1) [14], then

$$\begin{aligned}\frac{\partial^2 D_\phi(\theta_1,\theta_2)}{\partial\theta_{1i}\partial\theta_{1j}} &= \frac{\partial^2 D_\phi(\theta_1,\theta_2)}{\partial\theta_{2i}\partial\theta_{2j}} = I(\theta)\\ \frac{\partial^2 D_\phi(\theta_1,\theta_2)}{\partial\theta_{1i}\partial\theta_{2j}} &= -I(\theta)\end{aligned} \tag{57}$$

Using the arguments in (55)-(57) Equation (54) reduces to

$$\begin{aligned}&\frac{1}{2}\left(\widehat{\theta}_1-\theta_1\right)' I(\theta_1)\left(\widehat{\theta}_1-\theta_1\right)+\frac{1}{2}\left(\widehat{\theta}_2-\theta_1\right)' I(\theta_1)\left(\widehat{\theta}_2-\theta_1\right)\\ &-\left(\widehat{\theta}_1-\theta_1\right)' I(\theta_1)\left(\widehat{\theta}_2-\theta_2\right)+o\left(\left\|\widehat{\theta}_1-\theta_1\right\|^2\right)+o\left(\left\|\widehat{\theta}_2-\theta_2\right\|^2\right)\end{aligned} \tag{58}$$

Further,

$$\frac{2}{\phi''(1)}D_\phi\left(\widehat{\theta}_1,\widehat{\theta}_2\right)=\left(\widehat{\theta}_1-\theta_2\right)' I(\theta_1)\left(\widehat{\theta}_1-\theta_2\right)+o\left(\left\|\widehat{\theta}_1-\theta_1\right\|^2\right)+o\left(\left\|\widehat{\theta}_2-\theta_2\right\|^2\right) \tag{59}$$

Assuming that a change point τ divides the data into two heterogeneous parts with the parameters θ_1,θ_2 before and after the change point respectively with sample sizes τ, $(n-\tau)$ respectively, then by the regularity conditions the mles's are such that

$$\begin{aligned}\sqrt{\tau}\left(\widehat{\theta}_1-\theta_1\right) &\to N\left(0, I(\theta_1)^{-1}\right)\\ \sqrt{n-\tau}\left(\widehat{\theta}_2-\theta_2\right) &\to N\left(0, I(\theta_2)^{-1}\right)\end{aligned} \tag{60}$$

For $n\to\infty, \tau\to\infty, (n-\tau)\to\infty$

Let $\lambda=\frac{\tau}{n}\in(0,1)$ then,

$$\begin{aligned}\sqrt{\frac{\tau(n-\tau)}{n}}\left(\widehat{\theta}_1-\theta_1\right) &\to N\left(0, \lambda I(\theta_1)^{-1}\right)\\ \sqrt{\frac{\tau(n-\tau)}{n}}\left(\widehat{\theta}_2-\theta_2\right) &\to N\left(0, (1-\lambda) I(\theta_2)^{-1}\right)\end{aligned} \tag{61}$$

By the assumption of the null hypothesis $\theta_1=\theta_2=\theta_0$,

$$\sqrt{\frac{\tau(n-\tau)}{n}}\left(\widehat{\theta}_2-\widehat{\theta}_1\right)\to N\left(0, I(\theta_0)^{-1}\right) \tag{62}$$

under the assumption that the parameter estimates are consistent.

Suppose that under the maximum likelihood estimation for a sample of fixed size n, $\widehat{\theta}_n\to\theta$ as $n\to\infty$. By the law of large numbers, the observed informa-

tion matrix is such that,

$$I_n(\theta)=\left[\frac{1}{n}\sum_{i=1}^{n}-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x;\theta)\right]\to\left[E\left(-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x;\theta)\right)\right]=I(\theta) \quad (63)$$

If we substitute $\widehat{\theta}_n$ for θ

$$I_n(\theta)=\left[\frac{1}{n}\sum_{i=1}^{n}-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x;\theta)\right]_{\theta=\theta_n}\to\left[E\left(-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x;\theta)\right)\right]=I(\theta) \quad (64)$$

which is defined as a consistent estimator of the information matrix.

The elements of $I(\theta)$ are continuous in θ and it holds that

$$I_n(\theta)\to I(\theta) \text{ as } n\to\infty \quad (65)$$

From Equation (59) we obtain

$$\frac{\tau(n-\tau)}{n}\left(\widehat{\theta}_1-\theta_2\right)' I(\theta_1)\left(\widehat{\theta}_1-\theta_2\right)+o\left(\left\|\widehat{\theta}_1-\theta_1\right\|^2\right)+o\left(\left\|\widehat{\theta}_2-\theta_2\right\|^2\right) \quad (66)$$

From Equation (9) and Equations (56)-(66) then the test statistic can be expressed as

$$D_{n\tau}=\max_{\tau\in N(\epsilon)}\frac{\tau(n-\tau)}{n}\left\{\left(\widehat{\theta}_1-\theta_2\right)' I(\theta_1)\left(\widehat{\theta}_1-\theta_2\right)+o\left(\left\|\widehat{\theta}_1-\theta_1\right\|^2\right)+o\left(\left\|\widehat{\theta}_2-\theta_2\right\|^2\right)\right\} \quad (67)$$

Let

$$\begin{aligned}&\max_{\tau\in N(\epsilon)}\frac{\tau(n-\tau)}{n}\left\{\widehat{\left(\widehat{\theta}_1-\theta_2\right)' I(\theta_1)}\left(\widehat{\theta}_1-\theta_2\right)\right\}=W_{n\tau}\\&\max_{\tau\in N(\epsilon)}D_{n\tau}=\max_{\tau\in N(\epsilon)}W_{n\tau}+o\left(\left\|\widehat{\theta}_1-\theta_1\right\|^2\right)+o\left(\left\|\widehat{\theta}_2-\theta_2\right\|^2\right)\end{aligned} \quad (68)$$

But

$$o\left(\left\|\widehat{\theta}_1-\theta_1\right\|^2\right)=o_p(1)$$

$$o\left(\left\|\widehat{\theta}_2-\theta_2\right\|^2\right)=o_p(1)$$

Since the second and third terms of 67 are $o_p(1)$ then the distribution of $D_{n\tau}$ is similar to that of $W_{n\tau}$.