Spring 2022

Bootstrapping the Likelihood Ratio Test To Determine Change Points

Lili Donovan

Bootstrapping the Likelihood Ratio Test to Determine Change Points

Lili Donovan

June 2022

Advised by Kimihiro Noguchi and Ramadha Piyadi Gamage

Abstract

To detect mean changes in a sequence of independent observations from a certain distribution, the likelihood ratio (LR) test can be applied. Its p-value is typically computed using the asymptotic distribution of the test statistic, which may be unreliable when the number of observations is small. To overcome the problem, a simulation study is carried out to analyze the empirical Type I error rate and power of the LR test using the parametric and nonparametric bootstrap methods. Under both normal and exponential distributions, it is found that the nonparametric bootstrap makes the test robust for small, moderate, and large sample sizes.

1 Introduction to Bootstrap-Based Change Point Analysis

The basic idea of change point analysis is to detect structural changes such as mean, variance, distribution, etc., in a sequence of random observations. For example, consider an application to climate change, where a change in the mean or variance of the data set indicates a new trend in temperature or weather. In this case, identifying the location of change would pinpoint a date in which the new trend emerged. Change point analysis also has different medical applications, for instance with the annual flu, where trends in daily case counts can help estimate the beginning and end of the flu season. Another application can be found in quality control, where we suppose that we are in charge of monitoring some process to ensure that its observations stay within some range. That way, we may be able to minimize loss should we detect changes in the behavior of these observations.

There are several methods to identify change points and their location, such as maximum likelihood ratio test, information approach, Bayesian test, non-parametric test, and stochastic process. These methods determine if there is a change point in a sequence and provide an estimation of the change location. As there are often multiple change locations in a sequence, to find these locations, a combination of one of the the previous methods and the *binary segmentation procedure*, proposed by (Vostrikova, 1981), can be used. The binary segmentation procedure recursively finds change points by breaking up the sequence of random variables at a change point location. Then, we look for a new change point on the subsequent sub-sequences.

Let us now consider formulating the null and alternative hypothesis for the LR test. Using the binary segmentation procedure, the change point problem under the assumption that only one change point exists is equivalent to a hypothesis test with the null hypothesis:

$$H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_n, \tag{1}$$

and alternative hypothesis,

$$H_a \colon \mu_1 = \mu_2 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n, \tag{2}$$

for some $1 < k < n$, where k is the unknown change point location.

Traditionally, the statistical significance of a potential candidate change point in the change point analysis is computed using using the asymptotic distribution of the test statistic. However, the results may be unreliable when the number of observations is small. To improve the robustness of the change-point analysis, we present a simulation study which applies the normal-based parametric bootstrap and non-parametric bootstrap to the likelihood ratio (LR)

change point test statistic. The results show that the nonparametric bootstrap LR test is robust for both the normal and non-normal observations, including the exponential distribution.

Let $\{x_i\}_{i=1}^n$ be a sequence of independent and standardized normal random observations. Furthermore, let

$$S = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_k = \sum_{i=1}^{k}(x_i - \bar{x}_1)^2 + \sum_{i=k+1}^{n}(x_i - \bar{x}_2)^2.$$

The difference $S - S_k$ is mathematically equivalent to the LR statistic, which is discussed further in Section 3. Now, we wish to find k such that this difference is maximized, as this k corresponds to the most likely location of a change point. Thus, we take our test statistic to be

$$U = \max_{1<k<n} \sqrt{S - S_k}.$$

The asymptotic distribution of U is obtained after a slight transformation. Let

$$\mathcal{U} = a_n^{-1}(U - b_n),$$

where $a_n = (2\log\log n)^{-1/2}$ and $b_n = a_n^{-1} + \frac{1}{2}a_n \log\log\log n$. Then, $\mathcal{U}$ asymptotically has the extreme value distribution (Chen and Gupta, 2012), where the cumulative distribution function (cdf) is

$$F_{\mathcal{U}}(x) = e^{-2\pi^{1/2}e^{-x}}, -\infty < x < \infty.$$

In practice, $F_{\mathcal{U}}(x)$ gives an approximate p-value, and we reject H_0 when p-value $< \alpha$ for some desired significance level α. However, $F_{\mathcal{U}}(x)$ is less accurate for a small sample size, resulting in a less reliable p-value.

Therefore, the goal of using bootstrap is to make the test more robust; that is, the empirical Type I error rate α_e is made close enough to the desired significance level α so that the test decision is reliable. In this report, we explore two types of bootstrap methods to seek an appropriate test that achieves the goal. The first method is called *parametric bootstrap*, which creates resamples from an assumed distribution, such as the normal distribution. The other method is called *nonparametric bootstrap*, which creates resamples by sampling original observations with replacement. The nonparametric bootstrap offers a robust alternative to the large-sample approximation or parametric bootstrap for statistical inference and is valid for a wide range of data generating processes.

This report is organized as follows. First, we derive the maximum likelihood estimation of the parameters for the normal and exponential distributions, which are then used to construct the LR change point test statistics. Then, to evaluate the robustness of these tests, a comprehensive simulation study on the empirical

Type I error rate is conducted under various sample sizes and two distributions (normal and exponential) for the normal-based LR change point test statistic. Next, the normal-based LR test statistic is applied to detect change points from genome data on copy number variation to demonstrate the effectiveness of the nonparametric bootstrap approach.

2 Derivations of Maximum Likelihood Estimation (MLE)

The maximum likelihood estimation of the parameters associated with normal and exponential distributions are derived here. These parameters are then used in the derivation of the LR change point test statistic.

2.1 Normal Distribution

Assuming a known variance of $\sigma^2 = 1$, under each of H_0 and H_a, we calculate a maximum likelihood estimator (MLE) of μ.

To derive the MLE, we first must know about the likelihood function. This function is similar to a joint pdf, but instead of looking at it in terms of value(s) x_i that a distribution may take, the likelihood function is calculated in terms of the unknown parameters, which is μ in this case.

Under H_0, assuming that $X_1, \ldots, X_n$ are mutually independent with the marginal pdfs $f_{X_i}(x_i; \mu)$, $i = 1, \ldots, n$, the likelihood function is given by:

$$L_0(\mu) = f_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} f_{X_i}(x_i; \mu),$$

where we can substitute the pdf for the normal distribution $N(\mu, 1)$ into the equation to obtain:

$$L_0(\mu) = \frac{1}{(\sqrt{2\pi})^n} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2}.$$

The MLE comes from the maximizing the likelihood function and taking the derivative of the likelihood function with respect to μ. Typically, the log-likelihood function is used in place of the likelihood function as we can avoid some of the more complicated mathematical computation with exponents, while still obtaining the same result.

By taking the log of $L_0(\mu)$, the log-likelihood function of a random sample coming from $N(\mu, 1)$ is given by

$$\ell_0(\mu) = n \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{j=1}^{n}(x_j - \mu)^2.$$

Then, we find μ for which ℓ_0 is maximized by setting its derivative equal to 0. That is,

$$\frac{d}{d\mu}\ell_0(\mu) = \frac{d}{d\mu}\left(n\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^2\right) = \sum_{j=1}^{n}(x_j-\mu),$$

which can be simplified to $\sum_{j=1}^{n} x_j - n\mu$. If $\hat{\mu}$ is such that

$$\frac{d}{d\mu}\ell_0(\hat{\mu}) = 0,$$

the statement can only be true if:

$$\sum_{j=1}^{n} x_j - n\hat{\mu} = 0.$$

Thus, the MLE under H_0 is given by

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The log-likelihood function attains the maximum at this MLE of μ if the second derivative, with respect to μ, yields a negative value under the null hypothesis. The second derivative is given by:

$$\frac{d}{d\mu^2}\ell_0(\mu) = -n,$$

which results in a negative value when n is a positive integer. Hence $\hat{\mu}$ is the MLE of μ under the null hypothesis (1).

Under H_a (2), there is only one change point. Therefore, the corresponding likelihood function is given by:

$$L_1(\mu_1,\mu_n) = f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\mu_1,\mu_n) = \prod_{i=1}^{k} f_{X_i}(x_i;\mu_1)\prod_{i=k+1}^{n} f_{X_i}(x_i;\mu_n),$$

where μ_1 is the mean of the sample and μ_n is the alternate mean found at the change point. Which can be simplified to the following equation by substituting in the actual pdfs,

$$L_1(\mu_1,\mu_n) = \frac{1}{(\sqrt{2\pi})^n}e^{-(\sum_{i=1}^{k}(x_i-\mu_1)^2+\sum_{i=k+1}^{n}(x_i-\mu_n)^2)/2},$$

noting that $X_i \sim N(\mu_1,1)$ for $i = 1,\ldots,k$ and $X_i \sim N(\mu_n,1)$ for $i = k{+}1,\ldots,n$.

We calculate the MLE for μ_1 and μ_n by taking the same steps used for deriving the MLE for μ. The main difference is that the partial derivatives of the

log-likelihood function needs be taken with respect to μ_1 and μ_n.

The log-likelihood function under H_a is:

$$\ell_1(\mu_1, \mu_n) = n \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\left(\sum_{i=1}^{k}(x_i - \mu_1)^2 + \sum_{i=k+1}^{n}(x_i - \mu_n)^2\right).$$

From here, we take the partial derivatives of the log-likelihood function with respect to μ_1 and μ_n to find the MLE.

The partial derivatives with respect to μ_1 and μ_n are

$$\frac{\partial}{\partial \mu_1}\ell_1(\mu_1, \mu_n) = \frac{\partial}{\partial \mu_1}\left[n \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\left(\sum_{i=1}^{k}(x_i - \mu_1)^2 + \sum_{i=k+1}^{n}(x_i - \mu_n)^2\right)\right]$$

$$= \sum_{i=1}^{k}(x_i - \mu_1) = \sum_{i=1}^{k} x_i - k\mu_1,$$

and

$$\frac{\partial}{\partial \mu_n}\ell_1(\mu_1, \mu_n) = \frac{\partial}{\partial \mu_n}\left[n \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\left(\sum_{i=1}^{k}(x_i - \mu_1)^2 + \sum_{i=k+1}^{n}(x_i - \mu_n)^2\right)\right]$$

$$= \sum_{i=k+1}^{n}(x_i - \mu_n) = \sum_{i=k+1}^{n} x_i - (n-k)\mu_n$$

Setting these partial derivatives equal to zero will determine the MLE for μ_1 and μ_n:

$$\sum_{i=1}^{k} x_i - k\hat{\mu}_1 = 0,$$

$$\sum_{i=1}^{k} x_i = k\hat{\mu}_1,$$

$$\hat{\mu}_1 = \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i.$$

Similarly, for μ_n:

$$\sum_{i=k+1}^{n} x_i - (n-k)\hat{\mu}_n = 0,$$

$$\sum_{i=k+1}^{n} x_i = (n-k)\hat{\mu}_n,$$

$$\hat{\mu}_n = \bar{x}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^{n} x_i,$$

where n is the sample size and k is the location of the change point (where μ is shifted).

The log likelihood functions yield maximum under the alternate hypothesis if the Hessian Matrix is negative definite. We calculate:

$$\begin{bmatrix} f_{\mu_1^2} & f_{\mu_1 \mu_n} \\ f_{\mu_n \mu_1} & f_{\mu_n^2} \end{bmatrix}$$

For μ_1 and μ_n, the second partial derivatives are

$$\frac{\partial}{\partial \mu_1^2} \ell_1(\mu_1) = -k,$$

and

$$\frac{\partial}{\partial \mu_n^2} \ell_1(\mu_n) = -(n-k).$$

The other second partial derivatives $f_{\mu_1 \mu_n}, f_{\mu_n \mu_1}$ should be equivalent when entered into the Hessian matrix, and we have:

$$\frac{\partial}{\partial \mu_1} \frac{\partial}{\partial \mu_n} \ell_1(\mu_1, \mu_n) = \frac{\partial}{\partial \mu_1} \left(\sum_{i=k+1}^{n} x_i - (n-k)\mu_n \right) = 0$$

and

$$\frac{\partial}{\partial \mu_n} \frac{\partial}{\partial \mu_1} \ell_1(\mu_1, \mu_n) = \frac{\partial}{\partial \mu_n} \left(\sum_{i=1}^{k} x_i - k\mu_1 \right) = 0$$

The resulting Hessian matrix:

$$\begin{bmatrix} -k & 0 \\ 0 & -(n-k) \end{bmatrix}$$

is negative definite, confirming that our calculated μ_1 and μ_n are the MLEs under the alternate hypothesis, noting that n and k are both positive values and that $k < n$ is true.

2.2 Exponential Distribution

The method of using the likelihood ratio procedure to detect change points in exponentially distributed data is similar to that for a normal distribution in Section 2.1. In this set-up, the null hypothesis states that the observations are independent and that they are all coming from the same exponential distribution, with pdf equal to $f_{X_i}(x_i;\lambda) = \lambda e^{-\lambda x_i}$, $x_i > 0$, $\lambda > 0$. On the other hand, the alternative hypothesis claims that these independent observations are coming from two different exponential distributions with parameters λ_1 and λ_n. Specifically, there is k, $1 < k < n$, for which $f_{X_i}(x_i;\lambda_1) = \lambda_1 e^{-\lambda_1 x_i}$, $i \leq k$, and $f_{X_i}(x_i;\lambda_n) = \lambda_n e^{-\lambda_n x_i}$, $i > k$. Note that the parameters λ, λ_1, and λ_n are the rate parameters of the exponential distributions.

Because the likelihood function is based on the pdf, we can first look at the likelihood function under H_0:

$$L_0(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

As we did with the normal distribution, we maximize the log-likelihood function to find the MLE. Under H_0, the derivative of the log-likelihood function is taken with respect to λ.

$$\ell_0(\lambda) = \ln(\lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}) = \ln(\lambda^n) - \lambda \sum_{i=1}^{n} x_i = n[\ln(\lambda) - \lambda \bar{x}].$$

Then, we obtain the first derivative of $\ell_0(\lambda)$ with respect to λ. Specifically,

$$\frac{d}{d\lambda}\ell_0(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

By setting it equal to zero, we derive the MLE of λ. That is:

$$\frac{n}{\hat{\lambda}} = \sum_{i=1}^{n} x_i,$$

so that

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}.$$

Finally, the second derivative test is calculated to confirm that the MLE found above yields a maximum. The second derivative of the log-likelihood function with respect to λ is

$$\frac{d}{d\lambda^2}\ell_0(\lambda) = -\frac{n}{\lambda^2},$$

which is a negative value as both n and λ are positive; thus the test confirms that the MLE of λ under the null hypothesis is $\hat{\lambda}$.

Under H_a, there are two different λ values, namely, λ_1 and λ_n, to take into account when considering the likelihood function:

$$L_1(\lambda_1, \lambda_n) = \left(\lambda_1 e^{-\lambda_1 \sum_{i=1}^{k} x_i}\right) \left(\lambda_n^{n-k} e^{-\lambda_n \sum_{i=k+1}^{n} x_i}\right),$$

and log-likelihood function:

$$\begin{aligned}\ell_1(\lambda_1, \lambda_n) &= \ln(\lambda_1^k e^{-\lambda_1 \sum_{i=1}^{k} x_i}) + \ln(\lambda_n^{n-k} e^{-\lambda_n \sum_{i=k+1}^{n} x_i}) \\ &= \ln(\lambda_1^k) - \lambda_1 \sum_{i=1}^{k} x_i + \ln(\lambda_n^{n-k}) - \lambda_n \sum_{i=k+1}^{n} x_i \\ &= k[\ln(\lambda_1) - \lambda_1 \bar{x}_k] + (n-k)[\ln(\lambda_n) - \lambda_n \bar{x}_{n-k}].\end{aligned}$$

Under H_a, two partial derivatives of the log-likelihood function $\ell_1(\lambda_1, \lambda_n)$ are taken with respect to λ_1 and λ_n, respectively. The partial derivatives are given by:

$$\frac{\partial}{\partial \lambda_1} \ell_1(\lambda_1, \lambda_n) = \frac{k}{\lambda_1} - \sum_{i=1}^{k} x_i$$

and

$$\frac{\partial}{\partial \lambda_n} \ell_1(\lambda_1, \lambda_n) = \frac{n-k}{\lambda_n} - \sum_{i=k+1}^{n} x_i,$$

for λ_1 and λ_n, respectively. Then, the resulting MLEs are given by:

$$\hat{\lambda}_1 = \frac{k}{\sum_{i=1}^{k} x_i} = \frac{1}{\bar{x}_k}$$

and

$$\hat{\lambda}_n = \frac{n-k}{\sum_{i=k+1}^{n} x_i} = \frac{1}{\bar{x}_{n-k}}$$

by setting the partial derivatives equal to 0.

The second partial derivatives with respect to λ_1 or λ_n are,

$$\frac{\partial}{\partial \lambda_1^2} \ell_1(\lambda_1, \lambda_n) = -\frac{k}{\lambda_1^2},$$

and

$$\frac{\partial}{\partial \lambda_n^2} \ell_1(\lambda_1, \lambda_n) = -\frac{n-k}{\lambda_n^2},$$

respectively.

The second partial derivatives $f_{\lambda_1 \lambda_n}$ and $f_{\lambda_n \lambda_1}$ are equivalent. They can be found by using the partial derivatives found above:

$$\frac{\partial}{\partial \lambda_n} \left(\frac{k}{\lambda_1} - \sum_{i=1}^{k} x_i\right) = 0$$

and

$$\frac{\partial}{\partial \lambda_1}\left(\frac{n-k}{\lambda_n} - \sum_{i=k+1}^{n} x_i\right) = 0$$

Thus, the Hessian Matrix is:

$$\begin{bmatrix} f_{\lambda_1^2} & f_{\lambda_1\lambda_n} \\ f_{\lambda_n\lambda_1} & f_{\lambda_n^2} \end{bmatrix} = \begin{bmatrix} -\frac{k}{\lambda_1^2} & 0 \\ 0 & -\frac{n-k}{\lambda_n^2} \end{bmatrix}.$$

The matrix is negative-definite and so the maximum of the log likelihood function, the MLE, is confirmed to be $(\hat{\lambda}_1, \hat{\lambda}_n)$.

3 Likelihood Ratio Test Statistic Derivation

The LR test is used to assess the goodness-of-fit of two models. In the parametric approach, the observations are assumed to follow a known probability distribution. Instead of checking for multiple change points at the same time, the binary segmentation procedure is used here. With the binary segmentation method, it suffices to test the null hypothesis with no change versus the alternative hypothesis of one change. In LR test, the log-likelihood function is derived under H_0 and H_a. Let $\hat{\theta}$ be the MLE under H_0. Similarly, let $\hat{\theta}_1$ and $\hat{\theta}_n$ be the MLE under H_a. Then, for a fixed k, the LR test is:

$$\Lambda = \frac{L_0(\hat{\theta})}{L_1(\hat{\theta}_1, \hat{\theta}_n)}.$$

Instead of Λ, it is typical to consider its transformation

$$-2\ln\Lambda = -2[\ell_0(\hat{\theta}) - \ell_1(\hat{\theta}_1, \hat{\theta}_n)].$$

We show that $-2\ln\Lambda$ can be simplified even further for the normal and exponential distribution cases.

3.1 Normal Distribution

Recall that

$$\ell_0(\mu) = n\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{j=1}^{n}(x_j - \mu)^2$$

and

$$\ell_1(\mu_1, \mu_n) = n\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\left(\sum_{i=1}^{k}(x_i - \mu_1)^2 + \sum_{i=k+1}^{n}(x_i - \mu_n)^2\right).$$

Therefore,

$$\begin{aligned}
-2\ln\Lambda &= -2[\ell_0(\hat{\mu}) - \ell_1(\hat{\mu}_1, \hat{\mu}_n)] \\
&= -2\left[-\frac{1}{2}\sum_{j=1}^{n}(x_j - \hat{\mu})^2 + \frac{1}{2}\left(\sum_{i=1}^{k}(x_i - \hat{\mu}_1)^2 + \sum_{i=k+1}^{n}(x_i - \hat{\mu}_n)^2\right)\right] \\
&= \sum_{i=1}^{n}(x_i - \hat{\mu})^2 - \left[\sum_{i=1}^{k}(x_i - \hat{\mu}_k)^2 + \sum_{i=k+1}^{n}(x_i - \hat{\mu}_{n-k})^2\right] \\
&= S - S_k,
\end{aligned}$$

where

$$S = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

and

$$S_k = \sum_{i=1}^{k}(x_i - \bar{x}_k)^2 + \sum_{i=k+1}^{n}(x_i - \bar{x}_{n-k})^2. \tag{3}$$

Now, let

$$V_k = k(\bar{x}_k - \bar{x})^2 + (n-k)(\bar{x}_{n-k} - \bar{x})^2. \tag{4}$$

Then, we can show that $V_k = S - S_k$ holds by considering the first component of (3), which in turn makes the statement of V_k easier to visualize. Remember that we would like to maximize V_k in the process of determining a likely change location k. Now,

$$\begin{aligned}
S_k = \sum_{i=1}^{k}(x_i - \bar{x}_k)^2 &= \sum_{i=1}^{k}(x_i - \bar{x} + \bar{x} - \bar{x}_k)^2 \\
&= \sum_{i=1}^{k}(x_i - \bar{x})^2 + 2\sum_{i=1}^{k}(x_i - \bar{x})(\bar{x} - \bar{x}_k) - k(\bar{x} - \bar{x}_k)^2.
\end{aligned}$$

The second component of the equation can be rewritten as

$$2\sum_{i=1}^{k}(x_i - \bar{x})(\bar{x} - \bar{x}_k) = 2(\bar{x} - \bar{x}_k)\sum_{i=1}^{k}(x_i - \bar{x}) = 2(\bar{x} - \bar{x}_k)(k\bar{x}_k - k\bar{x}) = -2k(\bar{x} - \bar{x}_k)^2.$$

By plugging this back to (3) we get,

$$S_k = \sum_{i=1}^{k}(x_i - \bar{x})^2 - k(\bar{x} - \bar{x}_k)^2 + \sum_{i=k+1}^{n}(x_i - \bar{x})^2 - (n-k)(\bar{x} - \bar{x}_{n-k})^2$$

Now, substituting for S, we have,

$$S_k = S - k(\bar{x} - \bar{x}_k)^2 - (n-k)(\bar{x} - \bar{x}_{n-k})^2,$$

confirming that $V_k = S - S_k$.

Thus, the test statistic under the null hypothesis, denoted by U^2, is given by the maximum of V_k. Furthermore, we may write $U = \sqrt{U^2}$ in the following ways. Let

$$T_k = \sqrt{\frac{n}{k(n-k)}} \left[\sum_{i=1}^{k} (x_i - \bar{x}) \right].$$

Then, $V_k = T_k^2$, where

$$V_k = \frac{n}{k(n-k)} \left[\sum_{i=1}^{k} (x_i - \bar{x}) \right]^2.$$

Therefore, the test statistic U can be written as

$$U = \max_{1<k<n} \sqrt{V_k} = \max_{1<k<n} T_k.$$

3.2 Exponential Distribution

Recall that

$$\ell_0(\lambda) = n[\ln(\lambda) - \lambda \bar{x}]$$

and

$$\ell_1(\lambda_1, \lambda_n) = k[\ln(\lambda_1) - \lambda_1 \bar{x}_k] + (n-k)[\ln(\lambda_n) - \lambda_n \bar{x}_{n-k}].$$

Then, we have

$$\begin{aligned} -2\ln\Lambda &= -2[\ell_0(\hat{\lambda}) - \ell_1(\hat{\lambda}_1, \hat{\lambda}_n)] \\ &= -2\left[n[\ln(\hat{\lambda}) - \hat{\lambda}\bar{x}] - k[\ln(\hat{\lambda}_1) - \hat{\lambda}_1 \bar{x}_k] - (n-k)[\ln(\hat{\lambda}_n) - \hat{\lambda}_n \bar{x}_{n-k}] \right] \\ &= -2\left[n\ln(\bar{x}) - k\ln(\bar{x}_k) - (n-k)\ln(\bar{x}_{n-k}) \right] \end{aligned}$$

as the LR test statistic for the exponential distribution.

4 Bootstrap Algorithms

Two bootstrap methods are applied to the LR test statistic for the change point analysis to improve the robustness and power when compared to the one based on the asymptotic distribution, especially when the sample size is small. These two methods are called parametric and nonparametric bootstrap, where the main difference arises in how the resamples are generated. We describe the algorithm for each bootstrap method below.

Algorithm for the Parametric Bootstrap Method:

1. In the first step, we generate n independent standard normal ($N(0,1)$) random observations as a resample. The generated observations agrees with the null hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_n$, i.e., no changes in the process or parameters. The process is repeated B times to create B resamples, where B is a sufficiently large number.

2. Using B resamples, we compute B test statistics $\{U_1^\star, U_2^\star, ...U_B^\star\}$, where $U_b^\star$, $b = 1, 2, \ldots, B$, are the LR test statistics for the b-th resample.

3. Then, we calculate the empirical Type I error rate using a Monte Carlo simulation. First, we generate M samples of size n from some standardized distribution under H_0 (e.g., $N(0,1)$), each of which generates an observed LR test statistic.

4. Let $\{U_1, U_2, ...U_M\}$ be the set of observed LR test statistics from M Monte Carlo samples. The bootstrap p-value for U_m, $m = 1, 2, \ldots, M$, is given by $p_m = \#\{U_m < U_b^\star, b = 1, \ldots, B\}/B$.

The non-parametric bootstrap method, as used in the simulations with normal and exponential distributions, follows the algorithm above with some slight adjustments. Instead of generating resamples from a given/known distribution, the non-parametric bootstrap creates B resamples from (appropriately adjusted) original sample with replacement. Specifically, the original dataset is divided into two segments according to the estimated change-point location. Then, *each segment is mean-centered according to its respective sample mean.* That way, the mean-centered dataset respects the null hypothesis of no change point, and it then can be used to generate resamples.

To test the robustness of the two bootstrap methods, the empirical Type I error rate is calculated as an estimate of the actual type I error rate and power curves are created to determine where the change points can most accurately be identified.

The actual Type I error rate,

$$P(\text{Reject } H_0 \mid H_0 \text{ is true}),$$

is approximated by the empirical Type I error rate

$$\alpha_e = \#\{p_m < \alpha, m = 1, \ldots, M\}/M,$$

where M is the number of Monte Carlo samples generated of size n and p_m is the p-value for the m-th sample, $M = 1, 2, \ldots, M$.

Another important aspect of hypothesis testing is the power of the test. The power is the probability that the correct decision is made given that the alternate hypothesis is true. That is,

$$P(\text{Reject } H_0 \mid H_a \text{ is true}).$$

The power is also approximated using the same formula

$$\#\{p_m < \alpha, m = 1, \ldots, M\}/M.$$

In our simulation study, we investigate both the robustness and power of the parametric and nonparametric bootstrap-based test under small, medium, and large sample sizes. The data are generated from the standard normal ($N(0,1)$) and standard exponential (Exp(1)) distributions.

5 Simulation Study

We set up a simulation study to assess the robustness and power of the normal-based LR based on the two bootstrap methods (parametric and nonparametric). For both the parametric and nonparametric bootstrap simulations, we use the sample sizes of 30, 100, 200, and 500 generated from the standard normal and standard exponential distributions. Then, the empirical Type I error rate and power of the test are computed using Monte Carlo simulations to determine a robust bootstrap method.

5.1 Empirical Type I Error Rate

The empirical Type I error rate as an estimate of the actual Type I error rate, as described above, is expected to be close to our nominal Type I error rate (significance level) $\alpha = 0.05$ for each simulation if the test is robust.

	$N(0,1)$	Exp(1)
n	Empirical α	
30	0.046	0.097
100	0.040	0.087
200	0.061	0.106
500	0.056	0.104

Table 1: Parametric bootstrap simulation results with $B = 1000$ and $M = 1000$ at $\alpha = 0.05$.

When the nonparametric bootstrap is applied, the empirical Type I error rates approach our predetermined significance level of $\alpha = 0.05$ regardless of the distributions. On the other hand, when the parametric bootstrap is applied, the empirical Type I error rates do not seem to approach $\alpha = 0.05$ for

the standard exponential distribution even with $n = 500$. Thus, we find nonparametric bootstrap more robust than parametric bootstrap. In addition, for the nonparametric bootstrap, we run additional simulations with $n = 20$ and $n = 50$ to better understand the performance of the test when the sample size is small to moderate. Overall, the nonparametric bootstrap test starts performing reasonably well at around $n = 100$ for both the standard normal and standard exponential distributions.

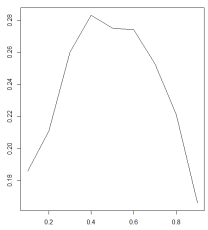
	$N(0,1)$	Exp(1)
n	Empirical α	
20	0.149	0.142
30	0.117	0.117
50	0.090	0.090
100	0.070	0.069
200	0.061	0.060
500	0.058	0.055

Table 2: Nonparametric bootstrap simulation results with $B = 1000$ and $M = 1000$ at $\alpha = 0.05$.

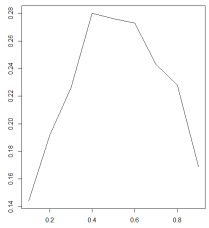
5.2 Empirical Power Curves

Using the nonparametric bootstrap, the following power curves are generated under several conditions. For each sample size, the change points are placed in several different relative locations; from 0.1 (at the first 10% of the data) to 0.9 (at the last 10% of the data) with 0.1 as the step size. The change of mean level that we identify is of step 0.5 for each distribution and each sample size. Given that the alternate hypothesis is true, a change point is present, implying that a more powerful test is shown by the one whose power is higher.

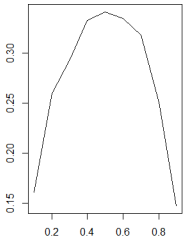
The following power curves indicate that the change point is more likely to be found when it is located in the middle of the dataset, where the power curve peaks. The power itself is improved when the sample size is increased; as can be seen in the increased power value for the sample sizes larger than 100. For the sample sizes smaller than 100, the peak power value still occurs in the middle of the dataset, but the value is lower than 1, indicating the change point is not always accurately found.



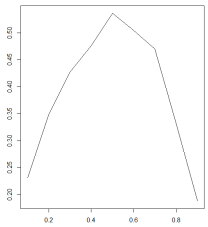
(a) Normal ($n = 20$)



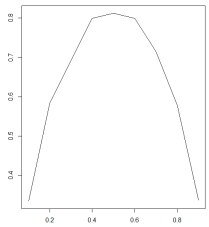
(b) Normal ($n = 30$)



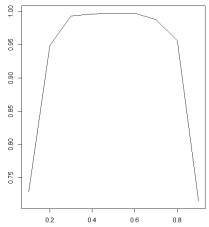
(c) Normal ($n = 50$)



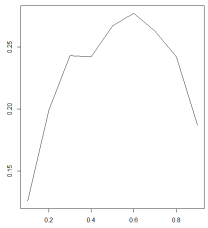
(d) Normal ($n = 100$)



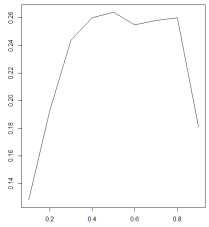
(e) Normal ($n = 200$)



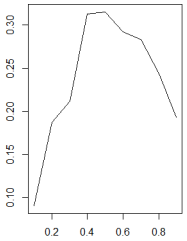
(f) Normal ($n = 500$)



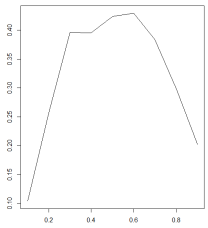
(g) Exp. ($n = 20$)



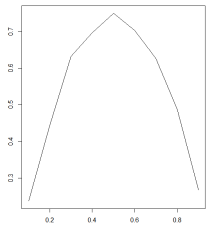
(h) Exp. ($n = 30$)



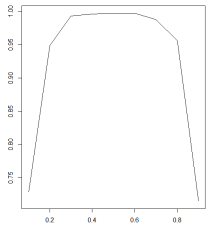
(i) Exp. ($n = 50$)



(j) Exp. n=100



(k) Exp. ($n = 200$)



(l) Exp. ($n = 500$)

Figure 1: Power Curves for Normal and Exponential Distributions. The relative locations and the empirical powers are on the x- and y-axis, respectively.

6 Real-Life Example

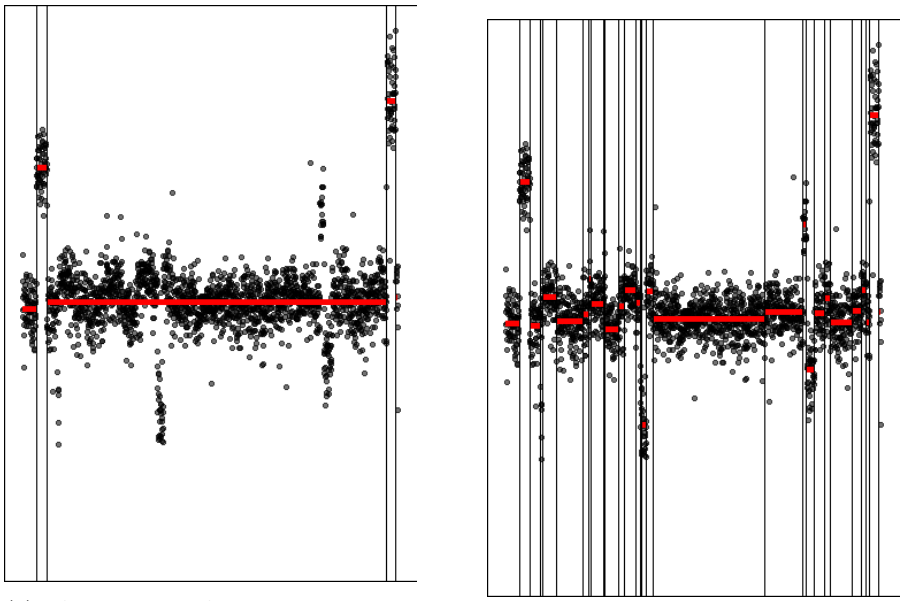
Determining where changes occur in DNA sequence requires the identification of copy number variation (CNV), which refers to the variation between individuals when considering the number of copies of a particular gene. This variation occurs when a particular gene is inserted or deleted. Variation is expected and is normal, considering each individual has their own unique gene sequence. However, it is of interest to medical professionals to know where and what variation occurs in an individual, to help in determining the risk associated with a certain variation.

Microarray-based comparative genomic hybridization (CGH) provided measures of DNA copy numbers and maps the values onto a sequence of genes (Chen and Wang, 2008). These measures are $\log_2$ mapping of the ratio comparing copy-number in the tested individual and the reference sample, compared at many positions along each chromosome. Insertions will cause the ratio to be larger than one, thus a positive $\log_2$ value. Deletions lead to a negative $\log_2$ since the ratio becomes less than one. There should be no zeros resulting from the variations. In practice, however, a zero is hard to avoid because of the difficulties of precise measurements. Furthermore, it is widely accepted that the resulting values follow a normal distribution with a mean of zero in the absence of variation. Thus, statistical analysis is needed to detect significant changes at possible change point locations.

For the real data application, we look at *The Fibroblast Cell Lines Data* (Snijders et al., 2001). This dataset provides results from a CGH experiment including $\log_2 T_i / \log_2 R_i$ measure with genomic location on multiple tissue cells (fibroblast cell lines) at 2,464 locations across the genome, where T_i and R_i are the copy numbers of the test and the reference samples, respectively, at location i. In our analysis, we compare the results based on the asymptotic distribution (Chen and Gupta, 2012) to the nonparametric bootstrap-based method. The nonparametric bootstrap-based is selected as our simulation confirms that it is more robust than the parametric bootstrap-based method when the empirical Type I error rates are compared under various sample sizes and distributions.

6.1 Results

In the following plots, the detected change points using the asymptotic distribution (left) (Prime, 2020) and nonparametric bootstrap (right) are displayed visually. The horizontal red line indicates the mean of each subsection of data. On the other hand, the vertical lines indicate that a change point is found at that location. At each vertical line, the dataset is split into two segments (the binary segmentation method). For each segment, the change point test statistic is applied again to detect more change point(s), if any, in these segments. It is clear that the nonparametric bootstrap is capable of detecting more possible change points while the asymptotic distribution-based approach seems to miss at least a few important change points.



(a) Asymptotic Analysis of Prime (2020)

(b) Nonparametric Bootstrap Analysis

Figure 2: Genome Analysis

Furthermore, the following is a table with data relevant to the genome data analysis. It includes the change point, the iteration (number of binary segmentations applied) in which the change point is detected, and the p-value associated with the respective change point.

Location	Iteration	p-value
75	3	0.000
134	2	0.000
194	4	0.000
201	3	0.003
281	6	0.000
426	7	0.004
458	9	0.008
470	10	0.007
538	11	0.002
548	8	0.000
624	9	0.000
652	5	0.000
715	6	0.000
717	8	0.000
718	7	0.031
740	4	0.000
746	6	0.038
774	5	0.000
810	6	0.001
1428	10	0.005
1636	9	0.000
1656	8	0.000
1700	9	0.000
1752	11	0.018
1786	10	0.000
1904	7	0.030
1959	9	0.000
1984	8	0.006
2003	1	0.000
2054	2	0.000

Table 3: Results with $B = 1000$ nonparametric bootstraps at $\alpha = 0.05$.

We notice that a majority of the change points are determined within the first ten iterations. Each of these locations is associated with a very low p-value, much less than the significance level of $\alpha = 0.05$.

7 Conclusion

Applying the nonparametric bootstrap to change point analysis is robust and powerful when the goal is to identify mean changes in normal and exponential distributions. Examining the empirical Type I error rate indicates that the change point is more accurately identified for a sample size of $n \geq 100$. The empirical power curves indicate a more accurate identification of a change point if the change is close to observation $n/2$ for a sample of size n, i.e., the middle of the dataset. These empirical power curves are also evidence that the larger the sample size ($n \geq 100$), the more accurate the analysis for identifying change points.

In the application to real data, the nonparametric bootstrap demonstrated improved change point identification when compared to the method based on the asymptotic distribution. Many more locations of change points were identified and their means were calculated, as shown in Figure 2. This happens due to the fact that the actual Type I error rate for the nonparametric bootstrap method is closer to the prespecified significance level α value when compared to the actual Type I error rate based on the asymptotic method of analysis, which tends to be more conservative. As a result, an increased number of change points identified by the nonparametric bootstrap method indicates a higher accuracy of detecting CNV in the genome sample. Improving detection accuracy is helpful in identifying chromosomal abnormalities.

When considering an additional medical application of change point analysis, we can look at functional magnetic resonance imaging (fMRI) as a technique for studying blood flow changes to the brain. When data from fMRI is acquired during a resting state, it can be difficult to determine the exact location of a deviation from stationary. In this 'rest' state, the brain is relatively free from external stimuli and so any changes in blood flow can be attributed to brain activity as a reaction to internal changes (Aston and Kirch, 2012). While these deviations can be hard to identify by eye, change point analysis can help analyze the fMRI plots developed at 'rest'.

References

Aston, J. A. and Kirch, C. (2012). Evaluating stationary via change-point alternatives with applications to fmri data. *The Annals of Applied Statistics*, 6(4).

Chen, J. and Gupta, A. K. (2012). Parametric statistical change point analysis: with applications to genetics, medicine, and finance.

Chen, J. and Wang, Y.-P. (2008). Evaluating stationary via change-point alternatives with applications to fmri data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):529–541.

Prime, N. (2020). Likelihood Ratio Approach for Detecting Mean Changes. Senior Project Report, Western Washington University.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., et al. (2001). Assembly of microarrays for genome-wide measurement of dna copy number. *Nature genetics*, 29(3):263–264.

Vostrikova, L. J. (1981). Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, **24**:55–59.