Likelihood Ratio Approach for Detecting Mean Changes

Noah Prime

Advised by Ramadha Piyadi Gamage

Abstract

The ability to detect changes in the structure of sequential data is practical in a wide variety of fields. In this paper, we take an in depth look at a likelihood ratio based method for detecting mean changes in a sequence of normally distributed observations described by Chen et al. (2012) and contrast it with a model selection based approach built around the Schwarz Information Criterion (SIC). These methods operate under the assumption of normality with homoscedastic and known variance. Simulations are carried out to compare the power and the empirical Type I error for likelihood ratio and Schwarz information criterion methods. A real data application has been performed to emphasise the validity and the importance of such analysis. While these methods are narrow in scope, the underlying ideas can be extended to apply change point analysis on sequential data of any type.

1 Introduction

1.1 Background

Perhaps we're in charge of monitoring some process which results in us having a sequence of repeated measures. It may be paramount that this measures stay within some range. We may be able to profit or at least minimize loss should we detect changes in the behavior of these measures. The basis of change point analysis is to detect changes in distribution (behavior) that these variables are coming from. That is, the random variables may come from a different distribution(s) as the sequence progresses.

The origins of these considerations comes from a quality control perspective, where it is of upmost importance that products be produced according to strict specifications from start to end of production. Then, one would want to find the point where products quality began to falter in order to potentially identify the source of the problem. Even better, identifying divergence from the desired distribution as soon as it happens. This is called *online* change point analysis which is referred to as "*Sequential Change Point Analysis*". There is also *offline* analysis which is when you have and analyze the whole of the data at once. In this paper we will refer to "offline change point analysis" just as "change point analysis".

In statistics the change point problem concerns both detecting whether or not a change or changes have occurred, and identifying the locations of any such changes. Since its beginnings, many more practical applications of change point analysis have been found. It is used to identify changes in the fluctuation of a stock price, quality control, to see if the lowering of speed limits resulted in a change in the traffic mortality rate, copy number changes in a DNA sequence, etc. (Chen et al. 2012), geology data analysis, etc. For example, the daily stock market records of the USA show that the stock price for any company fluctuates daily. Although the fluctuation of any stock price is normal according to the theory of economics, there are some shifts that are abnormal and worth investors' special attention. A change point problem in this scenario can be interpreted as whether a given incident happened in the world has caused a statistically significant change in the stock market of the USA.

1.2 Set-Up

Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables coming from respective distribution functions $F_1, F_2, ..., F_n$. If we allow for the possibility of q change points, where $q \geq 1$, the alternative hypothesis becomes:

$$H_a : F_1 = F_2 = \cdots = F_{k_1} \neq F_{k_1+1} = \cdots = F_{k_2} \neq F_{k_2+1} = \cdots \; F_{k_q} \neq F_{k_q+1} = \cdots = F_n$$

where $1 < k_1 < k_2 < \cdots < k_q < n$, are the q unknown locations of the change points.

There are several methods to identify and estimate the change points in the change point problem such as maximum likelihood ratio test, information approach, Bayesian test, non-parametric test, and stochastic process. The aforementioned methods allow one to determine if there is a change point in a sequence and provide an estimation of the location. To find multiple change locations in a sequence, which is often the case, a combination can be used of one of the the previous methods and the *binary segmentation procedure* proposed by Vostrikova (1981). This process recursively finds change points by breaking up the sequence of random variables at a change point location, then looking for a new change point on the subsequent sub-sequences. Explicitly, we start with a sequence of random variables $X_1, X_2, ..., X_n$ from distributions $F_1, F_2, ..., F_n$ respectively, and test the null hypothesis:

$$H_0 : F_1 = F_2 = \cdots = F_n,$$

vs the alternative hypothesis:

$$H_a : F_1 = F_2 = \cdots = F_k \neq F_{k+1} = F_{k+2} = \cdots = F_n.$$

If the null hypothesis is not rejected, then the process is stopped and we conclude that there is no change. If the null hypothesis is rejected, the sequence is split at k, and the same test is done on the resulting two sequences- before and after the change point found. This continues until no sub-sequences have change points. Then, the collection of change points found during this procedure gives the estimated total number and locations of change points in the original sequence.

This could very well be changed to a one sided test (for instance $F_k < F_{k+1}$) but only the two sided case will be covered here.

It is common that we consider a sequence of random variables which come form the same family of distributions. Then, we are testing whether the parameter(s) of the distribution change at some location(s) along the sequence. More specifically, in this paper we investigate methods used to find changes in the mean of a normal distribution assuming known and constant variance. That is, we assume $X_i \sim N(\mu_i, \sigma^2)$ and our null and alternative hypotheses become:

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

vs

$$H_a : \mu_1 = \mu_2 = \cdots = \mu_{k_1} \neq \mu_{k_1+1} = \cdots = \mu_{k_2} \neq \mu_{k_2+1} = \cdots\ \mu_{k_q} \neq \mu_{k_q+1} = \cdots = \mu_n \tag{2}$$

where q and $k_1, k_2, ..., k_q$ are unknown and have to be estimated.

Then, using the Binary Segmentation method, the change point problem under the assumption that only one change point exists is equivalent to a hypothesis test with null hypothesis,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$$

and alternative hypothesis,

$$H_a : \mu_1 = \mu_2 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n \tag{3}$$

for some $1 < k < n$ which is the unknown location of the change point.

2 Methods

A brief outline of the methods that are used in this paper to test the hypothesis whether a change has occurred or not for independent data is given below.

2.1 The Likelihood Ratio Approach

Likelihood ratio test (LRT) is used to asses 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,

$$H_a : \theta_1 = ... = \theta_k \neq \theta_{k+1} = ... = \theta_n.$$

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables from the normal distribution with mean μ_i and common variance σ^2. Here, we will use a likelihood ratio based test statistic to determine if there is significant evidence to reject the null hypothesis of no change ($H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$). Recall that we're assuming that the variance of each distribution is the same. Without loss of generality, we assume the common variance $\sigma^2 = 1$.

In LRT, the log-likelihood function is derived under H_0 and H_a.

Under H_0, the likelihood function is

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

and under H_a,

$$L_a(\mu_k, \mu_{n-k}) = \frac{1}{(\sqrt{2\pi})^n} e^{-(\sum_{i=1}^{k}(x_i-\mu_k)^2+\sum_{i=k+1}^{n}(x_i-\mu_{n-k})^2)/2}$$

where μ_k, μ_{n-k} are the population means before and after the change, respectively. Note that the maximum likelihood estimators for μ, μ_k, μ_{n-k} are

$$\hat{\mu} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\mu}_k = \bar{x}_1 = \frac{1}{k}\sum_{i=1}^{k} x_i, \qquad \hat{\mu}_{n-k} = \bar{x}_2 = \frac{1}{n-k}\sum_{i=k+1}^{n} x_i,$$

respectively. Let

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

For a fixed k, the log likelihood ratio statistic is

$$\begin{aligned}
-2\log\Lambda &= -2\log\frac{L_0(\hat{\mu})}{L_a(\hat{\mu}_k, \hat{\mu}_{n-k})} \\
&= -2\log\frac{e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2}}{e^{-(\sum_{i=1}^{k}(x_i-\mu_k)^2+\sum_{i=k+1}^{n}(x_i-\mu_{n-k})^2)/2}} \\
&= \sum_{i=1}^{n}(x_i - \hat{\mu})^2 - \sum_{i=1}^{k}(x_i - \hat{\mu}_k)^2 - \sum_{i=k+1}^{n}(x_i - \hat{\mu}_{n-k})^2 \\
&= S - S_k.
\end{aligned}$$

This difference $S - S_k$ tells us about the degree to which L_a is more likely than L_0. Note that this difference is guaranteed to be positive since $(\hat{\mu}_k, \hat{\mu}_{n-k})$ can explain as well as or better their respective sections of the sequence than μ can explain both. We want to find k such that this difference is maximized (which results in minimizing the likelihood ratio), as this k will correspond to the most likely location of a change point. Thus, we will take our test statistic to be

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

We will focus on the asymptotic distribution of a slight transformation of U. 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 $\mathcal{U}$ has cdf

$$F_{\mathcal{U}}(x) = e^{-2\pi^{1/2} e^{-x}}. \tag{4}$$

In practice then, we can compute the test statistic (U) and use the cdf of $\mathcal{U}$ to find a p-value, and reject the null hypothesis only if p-value $< \alpha$ for some desired significance level α. Rejecting the null hypothesis means accepting (3), that there is a change point in the sequence estimated to be at index k.

2.2 The SIC Approach to Mean Change

Change point problem can also be viewed from the aspect of the model selection. Schwarz Information Criterion (SIC) is used as a criterion for model selection, where the model with the lowest SIC value is considered to be the "best" model. SIC values will be calculated for the models under H_0 and H_a respectively. If the null model is selected then we conclude that there is no change. If the alternative model is selected then it will lead to the conclusion that there is a change. Simultaneously, the change location will be estimated.

Suppose that $x_1, x_2, ..., x_n$ is a sequence of independent and identically distributed random variables with probability density function $f(.|)$, where f is a model with t parameters, i.e.,

$$Model(t) : f(.|\tilde{\theta}) : \tilde{\theta} = (\theta_1, \theta_2, ..., \theta_t), \tilde{\theta} \in \Theta_t.$$

Here SIC value is used instead of AIC (Akaike Information Criterion) because the minimum AIC is not an asymptotically consistent estimator of the model order (Schwarz, 1978). Thus, SIC is expressed as,

$$SIC_k(t) = -2 \log L(\hat{\Theta}) + t \log n, t = 1, 2, ..., n$$

where $L(\hat{\Theta})$ is the maximum likelihood function of the model, t is the number of parameters to be estimated, n is the sample size and for our purposes the subscript k will denote the model associated with estimated change location k (the lack of such subscript will imply that there is no change, i.e., the model associated with H_0). SIC gives an asymptotically consistent estimate of the order of the true model.

Since the change location is unknown under the alternative hypothesis, SIC values will be calculated for all the possible values of k. Then the best model under the alternative hypothesis is the model associated with the smallest SIC value, i.e.,

$$SIC_{H_a} = \min_{1 \leq k \leq n} SIC_k(t).$$

According to the principle of the information criterion, H_0 is accepted if $SIC_{H_0} < SIC_{H_a}$, otherwise H_a is accepted for some k. This k is simultaneously the estimate of the location of the change point.

To test (1) vs (3), we compare the SIC for the two models associated with each hypothesis. Under H_0 we are estimating a single parameter (recall we're assuming there is known, common variance σ^2) μ with MLE $\hat{\mu} = \bar{x}_n$. The SIC becomes,

$$\begin{aligned} SIC(1) = SIC_{H_0} &= -2\log L(\hat{\mu}) + \log n \\ &= n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right) + \log n. \end{aligned}$$

Under the alternative hypothesis, we estimate two parameters μ_k and μ_{n-k}, the mean of the distribution up to the first change point and the mean of the distribution after the change point, with MLE's $\hat{\mu}_k = \bar{x}_k$ and $\hat{\mu}_{n-k} = \bar{x}_{n-k}$ where $\bar{x}_k$ is the sample mean of the first k observations in the sequence, and $\bar{x}_{n-k}$ is the sample mean of the last $n-k$ observations. The SIC (one per possible change location k) for these models then is,

$$\begin{aligned} SIC_k(2) &= -2\log L(\hat{\mu}_k, \hat{\mu}_{n-k}) + 2\log n \\ &= n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\left(\sum_{i=1}^{k}(x_i - \bar{x}_k)^2 + \sum_{i=k+1}^{n}(x_i - \bar{x}_{n-k})^2\right) + 2\log n. \end{aligned}$$

Chen et. al. (1997) recognizes the issue that perturbations in the data that do not indicate a change may still result in $SIC_{H_a} < SIC_{H_0}$ and therefor lead to a Type I error. In turn, it is proposed to view

$$\Delta = \min_{1 \leq k \leq n} [SIC_{H_a} - SIC_{H_0}]$$

as a test statistic. Similar to the LRT method which uses the asymptotic distribution of $\mathcal{U}$, a transformation of our test statistic U, we will the use the asymptotic null distribution of a transformation of Δ, $\dot{\mathit{\Delta}}$ to make inference. Let

$$\dot{\mathit{\Delta}} = a\log n - (2\log n - \Delta)^{1/2} - b\log n$$

where $a\log n = (2\log\log n)^{1/2}$ and $b\log n = 2\log\log n + \log\log\log n$. The asymptotic null distribution of $\dot{\Delta}$ is

$$F_{\dot{\mathit{\Delta}}}(x) = e^{-2e^{-x}}. \tag{5}$$

Note that

$$SIC_a - SIC_0 = \frac{1}{\sigma^2}[S_k - S] + \log n$$

where S_k, S are defined identically to their definitions in the Likelihood Ratio method.

3 Numerical Results

In this section, we conduct Monte Carlo simulations to investigate the performance of the LR approach and SIC approach in detecting mean changes in Normal distribution. We consider various sample sizes, and change locations as well.

3.1 Power

A simulation study is conducted to understand the power of the two methods described above. We look at sequences of length $25, 50$ and 100. For each sequence length, we generate a sequence of random observations of length k from $N(0, 1)$ and augment it with a sequence of random observations of length $n - k$ from $N(\mu_{n-k}, 1)$ where $\mu_{n-k} \in \{0.5, 1, 1.5, 2, 2.5\}$. For each sequence length there are seven values of k investigated, described by the floor of their percentage of the sequence length, i.e., the change point located at the 10-th percentage point would correspond with $k = 2, 5$ and 10 in the sequences of length $25, 50$ and 100 respectively. The floor is just to make sure we get an integer value. Each empirical power calculation is found via $10,000$ simulations with $\alpha = 0.05$ for both the LR and SIC methods.

Figure 1 illustrates the power of the two methods we have used against the change location as a percentage, for different values of μ_{n-k} and sequence length. Namely, the power of these tests are highly dependent on the location of the change point in the sequence. When the change location is centered in the sequence, these test have relatively high power even in short sequences with a large enough change in the mean. However, small changes in the location of the change point can drastically change the power of the tests in short sequences. Large sequences are more robust to these changes as shown by the flatter peaks of the curves.

That being said, for large changes in mean, these methods can be quite powerful, even in the short sequence case.

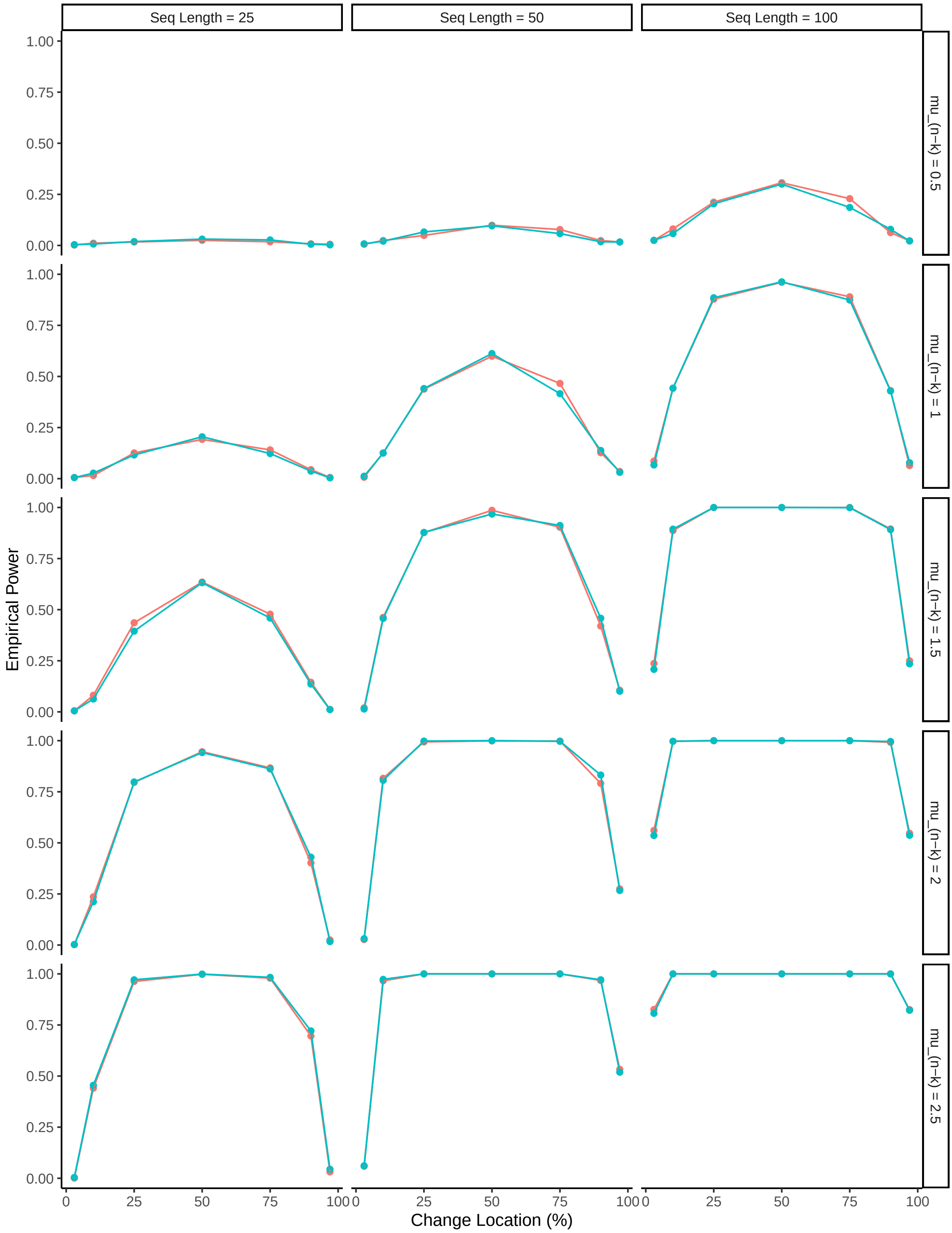


Figure 1: The empirical power curves of the LRT (blue) and SIC method (red) for detecting a change in the mean along a sequence of normally distributed, independent observations.

3.2 Empirical α

It is important to verify the validity of the two methods, particularly when dealing with asymptotic distributions. We want to have a gauge on how close the empirical significance level, α_e, i.e., the rate of incorrect rejections of H_0, is to our specified theoretical significance level, α.

Here, we perform a simulation with $100,000$ runs. In each trial we generate n random variables from the standard normal distribution and use the LR method to test for possible change points. Note that all n of the random observations were generated from the standard normal distribution. This is important to stress as it means that H_0 is true. Then, any rejections of H_0 will be classified as a Type I error. Since the LRT relies on asymptotics, we wanted to investigate α_e for differing sequence lengths ranging from quite short (25) to quite long $(5,000)$.

Table 1 demonstrates the conservative nature of the LRT for detecting change points, especially for short sequences. For example, the rate of Type I errors for sequences of length 25 was a tenth of what was desired. Even for quite long sequences α_e is a fraction of α. For example, sequences of length $5,000$ the LRT falsely rejected H_0 only about 1.45% of the time when the nominal $\alpha = 0.05$.

Sequence Length	α_e
25	0.00483
50	0.00570
75	0.00746
100	0.00808
200	0.00939
2000	0.01329
5000	0.01449

Table 1: α_e for different sample sizes when the nominal $\alpha = 0.05$.

In Table 2, for each sequence length the qth quantile value is given for the $100,000$ p-values generated. Clearly these values are much higher than the given quantile, especially for short sequences but again even for quite long sequences as well. One interpretation of this table can be demonstrated by looking at the 1st quantile when $n = 200$ which corresponds to 0.0512. This tells us that just 1% of p-values were less than 0.0512 when H_0 was true. Thus, when we use the LRT with $n = 200$ and $\alpha = 0.05 \approx 0.0512$, we are effectively testing at an approximately 99% significance level.

Sequence Length	1st Quantile	5th Quantile	10th Quantile
25	0.0691	0.1510	0.2141
50	0.0610	0.1383	0.2048
75	0.0578	0.1352	0.2007
100	0.0559	0.1315	0.1957
200	0.0512	0.1231	0.1874
2000	0.0417	0.1095	0.1722
5000	0.0414	0.1068	0.1683

Table 2: Quantile values of the $100,000$ p-values generated from the LRT when H_0 was true.

4 Application to CGH Data

Copy-number variation (CNV) refers to the variation of the number of copies of a particular gene between individuals. Variation occurs when there is an insertion or deletion of a particular gene sequence. Everyone has a unique combination of copy-numbers throughout their genome as it is immensely common for variation to occur. In fact, it turns out that it is quite common for large variation to occur, as in not just one deletion or insertion at a particular gene sequence, but maybe 5 copies at one location. The size and location of these variations is of particular interest as these are believed to be two of the most important clues in determining the risk associated with said variation. It is however no small task to find quantify copy-number variation in an individual in the first place.

Microarray-based comparative genomic hybridization (CGH) provides measures of DNA copy-number and maps values onto the genomic sequence. The given measures are the $\log_2$ mapping of the ratio of copy-number in the individual being tested and a reference sample at many positions along each chromosome. Deletions will lead to the ratio being less than one and thus a negative $\log_2$ and vice versa for an insertion. No variation should lead to a reading of zero. However, there are naturally difficulties in making such precise measurements. In practice, the measurements are actually a random variable widely accepted to follow a normal distribution and with mean of zero when not in the presence of variation. Consequently, statistical analysis is needed to detect significance at possible change point locations.

Here we take a look at the data used by Snijders et al. 2001 Snijders et al. (2001) provided by *The Fibroblast Cell Lines Data* 2009. The data provides the results from a CGH experiment including the $\log_2 T_i / \log_2 R_i$ measure with genomic location on multiple fibroblast cell lines (tissue cells) at over $2,000$ locations across the genome. We use both LR and SIC methods along with the binary segmentation procedure to analyze one of these cell lines in its entirety and on a chromosomal basis.

Both methods result in the same number and locations of change points. In the genome wide analysis there are four identified mean changes significant at the 0.05 significance level (Figure 2). By individual chromosomes, there are changes in chromosomes 1, 2, 7 (two changes), 16 and 23 (Figure 3). The genome wide analysis implies the existence in insertions of copies where we see the mean increase. The analysis by chromosome also is able to identify deletions on chromosome 7, 16 and 23. A medical professional may then use this information and possibly compare it with any known“key” locations, or locations in which changes have shown to be correlated with certain traits or diseases. The information learned from this test would also be crucial in any study attempting to understand copy number variation, for instance identifying those “key” locations. The discrepancy in the results between genome wide and by chromosome analysis is likely due to the limitations of the LRT including the lack of power of in certain situations which is discussed in the discussion section of this paper.

Note: R codes can be provided upon request.

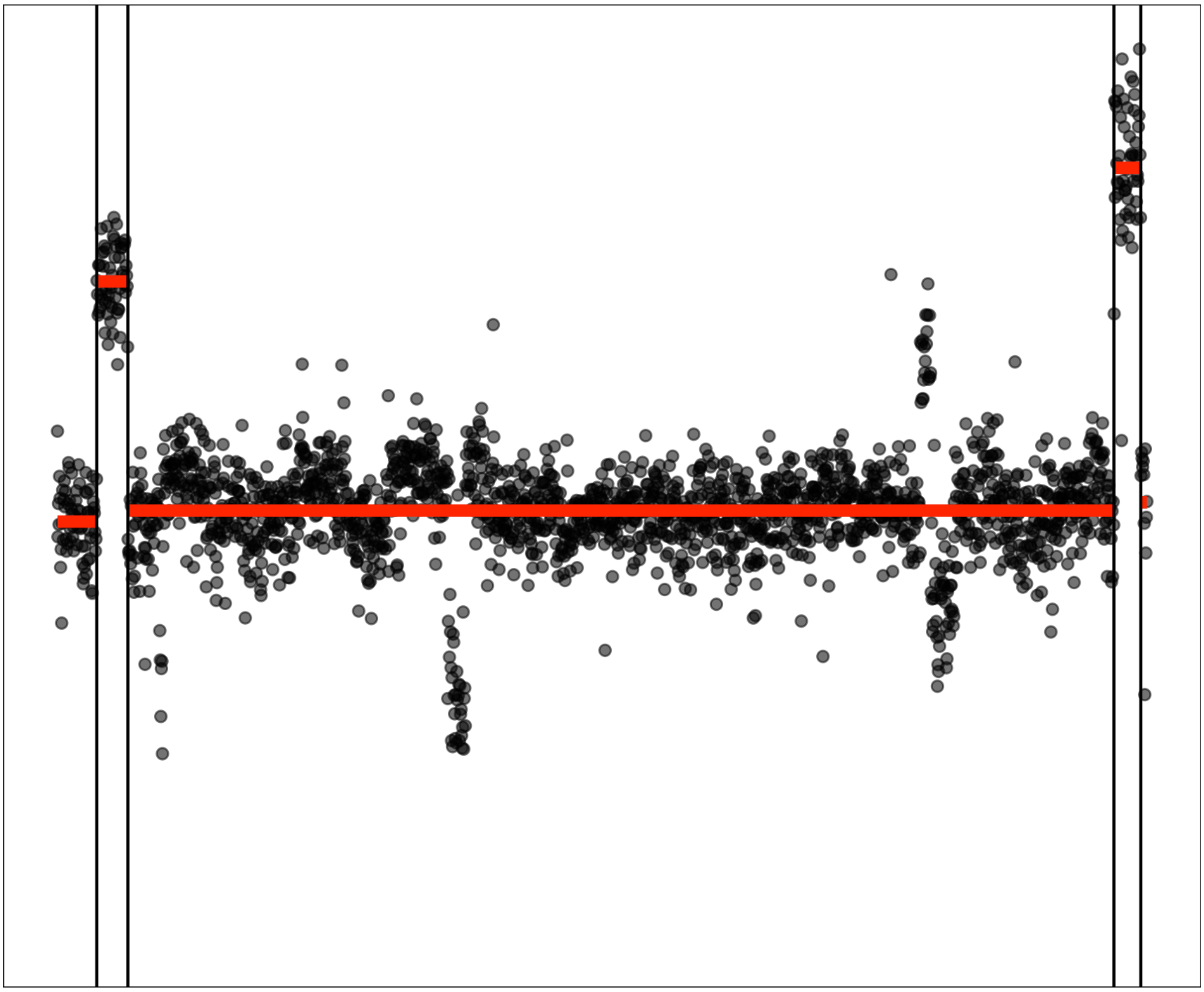
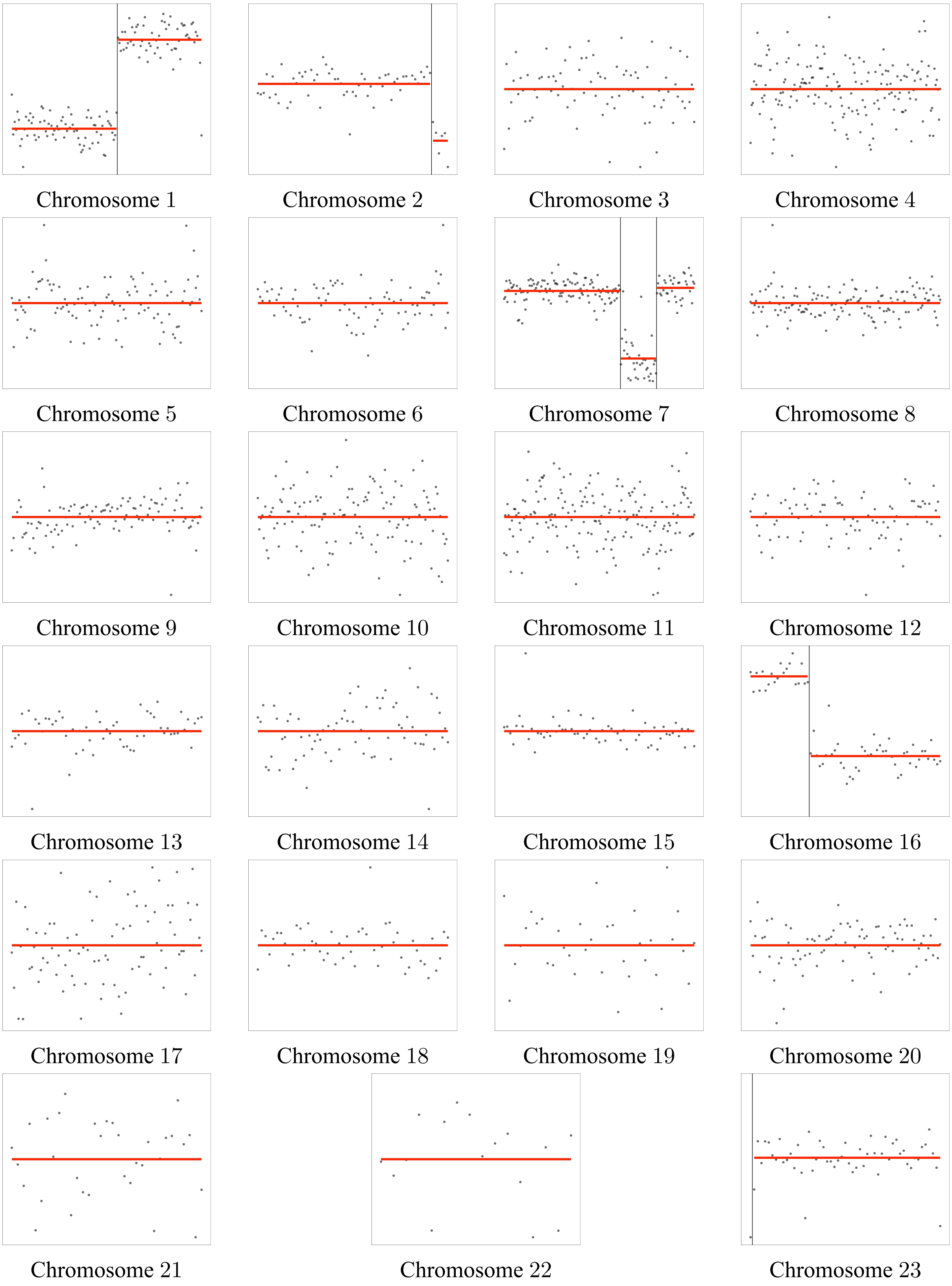


Figure 2: Change points as detected by the LRT on a genome wide search for copy number changes. Vertical lines at the change locations and horizontal lines drawn at the mean for each section.



Chromosome 1

Chromosome 2

Chromosome 3

Chromosome 4

Chromosome 5

Chromosome 6

Chromosome 7

Chromosome 8

Chromosome 9

Chromosome 10

Chromosome 11

Chromosome 12

Chromosome 13

Chromosome 14

Chromosome 15

Chromosome 16

Chromosome 17

Chromosome 18

Chromosome 19

Chromosome 20

Chromosome 21

Chromosome 22

Chromosome 23

Figure 3: Change point analysis by chromosome.

5 Discussion

In this paper, we analyze the LR and SIC methods for detecting change points in normally distributed data with known variance. Simulation study is used to describe the empirical power curves of these two methods and the empirical significance level of the LR method.

The two methods are quite similar in their test statistics and asymptotic null distributions of the test statistics. This leads to very similar performance in power of these test under all conditions investigated. On top of this, neither test has an advantage in computational complexity.

The empirical power curves reveal a major limitation of these methods, the drop off in performance based on the location of the change point in the sequence. The power to detect changes too early or too late in the sequence is dramatically lower than changes near the center. These methods are also not applicable to detecting a change at the first or last observation, as the MLE won't exist.

Consistent with other reports, convergence of the LRT statistic to the asymptotic distribution can be very slow. The test also tends to be quite conservative, especially for short sequences. This also can be partially attributed to the drop off in power for change locations at the edges of the sequence. To deal with these issues, multiple authors have proposed methods which involve a truncation of the sequence X_n or resampling and bootstrapping methods, and alternative methods to detect changes at the first or last observation. In this paper, we used our simulation study to provide empirical values to use for α to achieve a Type I error rate of $0.01, 0.05$ and 0.1.

In addition to the above limitations, there are particular qualities of a sequence which can decrease the power of these tests. For example, a sequence in which the mean is changed for a time but then returns to the original value. Since these methods on their own only detect single changes, the model associated with the alternative hypothesis will have a poor fit. In general, the performance of these methods are very dependent on the nature of the change(s) that don't have to do with the size of the change.

The Binary Segmentation Procedure is a simple and intuitive way for detecting multiple change points. This method has problematic limitations as well. The most obvious is the computational complexity that comes with iterative testing. Each change point detected requires three iterations of the test being used to detect a single change point. Another concern comes with the pairing of the BSP with the LR or SIC detection methods. As the BSP progresses, tests are being performed on sequences of decreasing length. As previously discussed, the power of these tests is dramatically smaller and thus can lead to change points going undetected due to the order in which the changes are detected.

The CGH data analyzed in this paper has unknown variance, but appears to satisfy the common variance assumption. Then, after standardization we are able to apply the method and the results validate the methods use, identifying reasonable change locations. However, the results are perhaps underwhelming due to the discussed limitations of the method. Looking at the chromosome wide analysis, one might expect more changes to be detected. This is further shown by the analysis done at individual chromosomes, as there in total more change locations detected. A limitation that was discussed that may have play a large role in this analysis is the sequence diverging and then returning to baseline. This appears to happen multiple times along the sequence which went undetected, but as you look at sections of the sequence, some of these are able to be detected since the sequence solely due to the way the sequence is split up.

6 Conclusion/Future Work

This paper demonstrates the ability of the LRT to detect changes in the mean of univariate normal data with known and common variance. This validates the use of the LRT method and more generally the practicality of change point analysis. Through the success of the application of the LRT to a real world data set (CGH data) we see the benefit in continuing exploration in the field.

The LRT is however far from perfect. Much work should be done to minimize the limitations of the method previously discussed. Alternatively, it may be a better use of resources to develop methods which don't suffer from the same limitations, particularly for the case where there may be more than one change location. While the pairing of the LRT with the BSP provides a simple way for detecting multiple changes, the LRT at each step operates blind to the fact that multiple changes may exist. For this reason, although such methods may be more difficult to come by, it may be advantageous to develop methods which can detect multiple changes simultaneously. This could circumnavigate the fluctuation in usefulness of the LRT with respect to the structure of the data.

Of course, change points are not unique to normal data. Methods to deal with data coming from many other distributions and non-parametric methods exist and deserve similar analysis. This will make a number of new data sets available to analyze from the change point perspective.

7 Appendix (Theoretical Results)

Deriving the Asymptotic Null Distribution of U

In this section, we will consider a sequence of random variables (X_n) with $X_i \sim N(\mu_i, \sigma_i^2)$. We are interested in testing the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$$

vs the alternative:

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

for some k. Let us recall the definition of our test statistic U.

Let

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

and

$$V_k = S - S_k.$$

Then, S is the sum of squared deviations under H_0, S_k the sum of squared deviations under H_a and V_k the difference between the two. The larger the difference between the two, the stronger the case for H_a, so we take $V_{k*} = \max_{1 \leq k \leq n-1} V_k = U^2$ as our test statistic.

An alternative but equivalent test statistics comes from the representation of V_k as

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

and letting

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

Then, $V_k = T_k^2$ or $|T_k| = \sqrt{V_k}$ so we can use our test statistic to be

$$U = \sqrt{V_{k*}} = \max_{1 \leq k \leq n-1} V_k = \max_{1 \leq k \leq n-1} |T_k|.$$

This is how our test statistic is defined and now we must consider the probability density function of U.

In this section we will provide the proof of two theorems which rely on serveral lemmas proved in Chen et al. 2012.

Lemma 2.2: $\{T_1, T_2, ..., T_{n-1}\}$ is a *Markov process.*

Definition: **Markov Process** – A random process where the probability of a future event depends only on the state attained in the previous event.

Theorem 2.1: The probability density distribution of the test statistic U is given by

$$f_U(x) = 2\Phi(x, 0, 1) \sum_{k=1}^{n-1} g_k(x, x) g_{n-k}(x, x),$$

where $\Phi(x, 0, 1)$ is the pdf of the standard normal distribution, $g_1(x, s) = 1$ for $x, s \geq 0$, and

$$g_k(x, s) = P(|T_i| < s, i = 1, ..., k-1 | |T_k| = x), x, s \geq 0.$$

Proof. We use the cdf $F_U(x)$ and consider $F_U(x + dx) - F_U(x)$. Note that we will let $dx \to 0$ making $F_U(x + dx) - F_U(x) = f_U(x)$. Then, $F_U(x + dx) - F_U(x)$ is the same thing as the probability that U is in the interval $(x, x + dx)$ and since $U = \max_{1 \leq k \leq n-1} |T_k|$ this is equivalent to

$$P \left\{ \bigcup_{i=1}^{n-1} [|T_i| \in (x, x + dx)] \bigcap [|T_i| > |T_j|, j \neq i] \right\}$$

or the probability that $|T_i|$ is in that interval and is the maximum of all $|T_k|'s$ for some i. Breaking up the last condition into $j < i$ and $j > i$ we get

$$P \left\{ \bigcup_{i=1}^{n-1} [|T_i| \in (x, x + dx)] \bigcap [|T_i| > |T_j|, j < i] \bigcap [|T_i| > |T_j|, j > i] \right\}.$$

Since if $|T_i|$ is the max then $|T_j|$ is not the max unless $i = j$, thus these sets are disjoint, so we can break this up into a sum,

$$\sum_{k=1}^{n-1} P\Big\{[|T_i| \in (x, x+dx)] \bigcap [|T_i| > |T_j|, j < i] \bigcap [|T_i| > |T_j|, j > i]\Big\}.$$

Calling the three sets separated by $\cap$, A, B, C respectively, we get

$$\sum_{k=1}^{n-1} P\{A \cap B \cap C\} = \sum_{k=1}^{n-1} P[A]P[B|A]P[C|AB].$$

Note that $\{T_1, T_2, ..., T_{n-1}\}$ is a Markov process (**Lemma 2.2**) which then allows us to state that $\{T_1, T_2, ..., T_{k-1}\}$ and $\{T_{k+1}, T_{k+2}, ..., T_{n-1}\}$ are independent given that $T_k = x$. We then can use this result to see that B, C are independent so $P[C|AB]$ simplifies to $P[C|A]$. Since $T_k \sim N(0, 1)$,

$$P(A) = 2\Phi(x, 0, 1) + o(dx)$$

(2 because of the absolute value).

$$P[B|A] = P[|T_j| < x, i = 1, ..., k-1 | |T_k| = x] + o(dx)$$

which is how $g_k(x, x)$ is defined plus $o(dx)$, and similarly,

$$P[C|A] = g_{n-k}(x, x) + o(dx).$$

Thus,

$$F_U(x + dx) - F_U(x) = \sum_{k=1}^{n-1} 2\Phi(x, 0, 1) g_k(x, x) g_{n-k}(x, x) + o(dx),$$

or with $dx \to 0$,

$$f_U(x) = \sum_{k=1}^{n-1} 2\Phi(x, 0, 1) g_k(x, x) g_{n-k}(x, x).$$

This completes the proof of the theorem. □

Then, though derived, the null distribution of U is very messy and difficult to work with. This is why we take

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

where $a_n = (2 \log\log n)^{-1/2}$ and $b_n = a_n^{-1} + \frac{1}{2} a_n \log\log\log n$, as this will turn out to have a much nicer and well known Gumble distribution.

Lemma 2.6:

$$\max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} = o_p(a_n).$$

I.e., the two statistics, $\max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}}$, $\max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}}$, which deal with the beginning of the sequence up to $n/\log n$ are asymptotically identical.

Lemma 2.7:

$$\max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} = O_p((\log\log\log n)^{1/2}).$$

I.e., the statistic dealing with the sequence up to $n/\log n$ has the same order as $O_p((\log\log\log n)^{1/2})$.

Lemma 2.8: For $-\infty < x < \infty$,

$$\lim_{n\to\infty} P\left[a_n^{-1}\left(\max_{1\le nt\le n/\log n}\frac{|B(t)|}{\sqrt{t}} - b_n\right)\le x\right] = \text{exp}\{-\pi^{1/2}e^{-x}\}$$

Lemma 2.9: The following holds as $n\to\infty$:

(i)

$$\max_{1\le nt\le[n/2]}\frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1\le nt\le[n/\log n]}\frac{|B(t)|}{\sqrt{t}} = o_p(a_n).$$

(ii)

$$\max_{1\le n(1-t)\le[n/2]}\frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1\le n(1-t)\le[n/\log n]}\frac{|B(t)-B(1)|}{\sqrt{t}} = o_p(a_n).$$

Theorem 2.5

$$\lim_{n\to\infty} P[a_n^{-1}(U-b_n)\le x] = \text{exp}\{-2\pi^{1/2}e^{-x}\}$$

Proof. Note that Yao and Davis (1986) show that our test statistic U is identiacally distributed as

$$\max_{nt=1,\ldots,n-1}|B_0(t)|/[t(1-t)]^{1/2}$$

where $t=k/n$, $B(t), 0 <\le t < \infty$ is a standard Brownian motion and $B_0(t) = B(t) - tB(1)$ is the Brownian Bridge. Then,

$$\begin{aligned}
&P[a_n^{-1}(U-b_n)\le x|H_0]\\
&= P[U\le a_nx+b_n]\\
&= P\left[\max_{1\le nt\le n-1}\frac{|B_0(t)|}{\sqrt{t(1-t)}}\le a_nx+b_n\right]\\
&= P\left[\max_{1\le nt\le n/2}\frac{|B_0(t)|}{\sqrt{t(1-t)}}\le a_nx+b_n, \max_{1\le n(1-t)\le n/2}\frac{|B_0(t)|}{\sqrt{t(1-t)}}\le a_nx+b_n\right]\\
&= P\left[\max_{1\le nt\le n/\log n}\frac{|B(t)|}{\sqrt{t}}\le a_nx+b_n, \max_{1\le n(1-t)\le n/\log n}\frac{|B(t)-B(1)|}{\sqrt{1-t}}\le a_nx+b_n\right]+o_p(1) \qquad (6)\\
&= P\left[\max_{1\le nt\le n/\log n}\frac{|B(t)|}{\sqrt{t}}\le a_nx+b_n\right]\cdot P\left[\max_{1\le n(1-t)\le n/\log n}\frac{|B(t)-B(1)|}{\sqrt{1-t}}\le a_nx+b_n\right]+o_p(1)\\
&\overset{n\to\infty}{\to} \text{exp}(-\pi^{-1/2}e^{-x})\cdot\text{exp}(-\pi^{-1/2}e^{-x}) \qquad (7)\\
&= \text{exp}(-2\pi^{-1/2}e^{-x}).
\end{aligned}$$

(6) from Lemma 2.9 and (7) from Lemma 2.8 □

References

Chen, J. and A. K. Gupta (1997). “Testing and Locating Variance Change Points with Application to Stock Prices”. In: *J. Am. Statistical Assoc.* **92**, pp. 739–747.

Chen, J. and A. K. Gupta (2012). *Parametric Statistical Change Point Analysis*. Boston: Birkhäuser.

Schwarz, G. (1978). “Estimating the Dimension of a Model”. In: *Annals of Statistics* **6**, pp. 461–464.

Snijders, A., N. Nowak, and R. et al. Segraves (2001). “Assembly of microarrays for genome-wide measurement of DNA copy number”. In: *Nature Genetics* **29**, pp. 263–264.

The Fibroblast Cell Lines Data (2009). `http://www.nature.com/ng/journal/v29/n3/full/ng754.html`.

Vostrikova, L. J. (1981). “Detecting ‘Disorder’ in Multidimensional Random Processes”. In: *Soviet Mathematics Doklady* **24**, pp. 55–59.