CONSISTENCY RESULTS IN MULTIPLE CHANGE-POINT PROBLEMS

by

Ennapadam Seshan Venkatraman

TECHNICAL REPORT NO. 24
March 1992

Prepared Under Grant MDA 904-89-H-2040
For The National Security Agency

Reproduction in Whole or in Part is Permitted
for any Purpose of the United States Government

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Consistency Results in Multiple Change-point Problems

by

Ennapadam Seshan Venkatraman

TECHNICAL REPORT NO. 24
March 1992

PREPARED UNDER GRANT MDA 904-89-H-2040
FOR THE NATIONAL SECURITY AGENCY

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Abstract

We will consider the problem of estimating the number and the locations of change-points in a process that has been observed. A related problem of interest is the estimation of the distribution functions between change-points.

First we will deal with the special case of a Gaussian process where only the mean of the process changes. The binary segmentation procedure and a procedure using the Schwarz criterion have been proposed for this problem. We will prove consistency results for these two procedures under weaker conditions on the number and the locations of the change-points. We will give an example where the binary segmentation procedure fails asymptotically while the other procedures are still consistent.

We will introduce the pseudo-sequential procedures for the general problem and give a prototype of the consistency statement and the method of proof for this class. These are sequential schemes for online detection of a change which have been adapted to our fixed sample problem. For the Gaussian case we will study a Bayesian method and a likelihood ratio method and will prove consistency theorems for both these types of procedures.

We will use Monte Carlo methods to compare the performance of these procedures in estimating the number of change-points, their locations and the mean function. We will conclude from the Monte Carlo results and the consistency theorems that the pseudo-sequential procedures, properly modified to take care of shooting over the change-point, performs as well as the others. We will derive some tail probability
approximations for some of the statistics used to compute the critical levels required for the Monte Carlo study.

Another problem we will consider is the exponential process with changes in the rate. We will study the Bayesian method for this case and prove that it is consistent.
Acknowledgements

I would like to thank my advisor Prof. David Siegmund for all the interest he showed in me and his constant support. I would also like to thank Prof. Richard Olshen for reading my thesis as well as helping me in general. My thanks also to Prof. Tze Lai for reading my thesis. Finally, I would like to thank all the others who helped me during my stay at Stanford.
Contents

Abstract iii

Acknowledgements v

1 Introduction 1
    1.1 Statement of the Problem 2
    1.2 Review 3
    1.3 Outline 4

2 Binary Segmentation 8
    2.1 Description of the Procedure 8
    2.2 Consistency Result 10
    2.3 A Counter Example 23

3 Schwarz Criterion 26
    3.1 Description of the Procedure 26
    3.2 Consistency Result 28

4 Pseudo-Sequential Procedures 37
    4.1 Description of the Procedure 37
    4.2 General Consistency Statement 38
    4.3 Special Cases 41
List of Tables

7.1 Procedures and Notations. .............................................. 69
7.2 (a) Mean functions $f_1$, $f_2$ and $f_3$. .......................... 69
7.2 (b) Mean functions $f_4$ and $f_5$. ................................. 69
7.2 (c) Mean function $f_6$. ............................................. 69
7.3 Frequency distribution of the estimated number of change-points for the mean function $f_1$. ........................................... 70
7.4 Frequency distribution of the estimated number of change-points for the mean function $f_2$. ........................................... 71
7.5 Frequency distribution of the estimated number of change-points for the mean function $f_3$. ........................................... 71
7.6 Frequency distribution of the estimated number of change-points for the mean function $f_4$. ........................................... 71
7.7 Frequency distribution of the estimated number of change-points for the mean function $f_5$. ........................................... 72
7.8 (a) Frequency distribution of the estimated number of change-points for the mean function $f_6$. ........................................... 72
7.8 (b) Frequencies in the range 0 - 4 for the binary segmentation procedure. 72
7.9 Mean absolute deviations of the estimated number of change-points from the true number. ................................................. 73
7.10 Simulation results for the counter example. .......................... 87
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.11</td>
<td>The four factors and the two levels of each of the factors we use for the experiment.</td>
<td>89</td>
</tr>
<tr>
<td>7.12</td>
<td>Factorial Experiment - Simulation results for the Bayesian method with $\delta = 1$.</td>
<td>90</td>
</tr>
<tr>
<td>7.13</td>
<td>Factorial Experiment - Simulation results for the mixture version of the Bayesian method.</td>
<td>91</td>
</tr>
<tr>
<td>7.14</td>
<td>Factorial Experiment - Simulation results for the generalized likelihood ratio method.</td>
<td>92</td>
</tr>
<tr>
<td>7.15</td>
<td>Factorial Experiment - Simulation results for the binary segmentation procedure.</td>
<td>93</td>
</tr>
<tr>
<td>7.16</td>
<td>Factorial Experiment - The main effects and the interactions for the four procedures.</td>
<td>94</td>
</tr>
<tr>
<td>7.17</td>
<td>The Problem of Nile - The minimal sum of squares and Schwarz criterion for both the original data and the transformed data.</td>
<td>100</td>
</tr>
<tr>
<td>7.18</td>
<td>The Problem of Nile - The estimated number and the locations of the change-points for all the procedures used.</td>
<td>100</td>
</tr>
<tr>
<td>A.1</td>
<td>Probability from Monte Carlo simulation and the approximation.</td>
<td>124</td>
</tr>
</tbody>
</table>
List of Figures

7.1 (a) The histograms of IMSE and PMAD for the mean functions $f_1$ - $f_3$ for the binary segmentation procedure. ............................................. 74

7.1 (b) The histograms of IMSE and PMAD for the mean functions $f_4$ - $f_6$ for the binary segmentation procedure. ............................................. 75

7.2 The histograms of IMSE and PMAD for the mean functions $f_1$ - $f_3$ for the procedure based on the Schwarz criterion. ................................. 76

7.3 (a) The histograms of IMSE and PMAD for the mean functions $f_1$ - $f_3$ for the mixture Bayesian procedure. ............................................. 77

7.3 (b) The histograms of IMSE and PMAD for the mean functions $f_4$ - $f_6$ for the mixture Bayesian procedure. ............................................. 78

7.4 (a) The histograms of IMSE and PMAD for the mean functions $f_1$ - $f_3$ for the generalized likelihood ratio procedure. ............................... 79

7.4 (b) The histograms of IMSE and PMAD for the mean functions $f_4$ - $f_6$ for the generalized likelihood ratio procedure. ............................... 80

7.5 (a) The histograms of IMSE and PMAD for the mean functions $f_1$ - $f_3$ for the fixed $\delta$ Bayesian procedure. ............................................. 81

7.5 (b) The histograms of IMSE and PMAD for the mean functions $f_4$ - $f_6$ for the fixed $\delta$ Bayesian procedure. ............................................. 82

7.6 (a) The histograms of IMSE and PMAD for the mean functions $f_1$ - $f_3$ for the first modified Bayesian procedure. ............................................. 83
7.6 (b) The histograms of IMSE and PMAD for the mean functions $f_4$ -
$f_6$ for the first modified Bayesian procedure. .................. 84
7.7 (a) The histograms of IMSE and PMAD for the mean functions $f_1$ -
$f_3$ for the second modified Bayesian procedure. ............... 85
7.7 (b) The histograms of IMSE and PMAD for the mean functions $f_4$ -
$f_6$ for the second modified Bayesian procedure. ............... 86
7.8 Q-Q plot of the effects for Bayesian method - $\delta = 1$. .......... 95
7.9 Q-Q plot of the effects for Bayesian method - Mixture. ........... 96
7.10 Q-Q plot of the effects for Likelihood Ratio method. ............ 97
7.11 Q-Q plot of the effects for Binary Segmentation Procedure. .... 98
7.12 (a) The Problem of Nile - Data and the likelihood function for the
binary segmentation procedure. .................. 101
7.12 (b) The Problem of Nile - The likelihood functions for the Bayesian
mixture method and the generalized likelihood ratio method. .... 102
7.12 (c) The Problem of Nile - The likelihood function for the fixed $\delta$
Bayesian method and the second modified mixture Bayesian method. 103
7.12 (d) The Problem of Nile - The likelihood function for the two segments
from the binary segmentation procedure and the data until the value
of the stopping time for the modified mixture Bayesian methods. ... 104
8.1 The British coal-mining disaster data and the Bayesian likelihood func-
tion to detect a change in the rate. .................. 113
Chapter 1

Introduction

A change-point is a point at which the stochastic behavior of a process changes, for example, it is the point at which an industrial process, which is in control, goes out of control. More formally, let $X_1, X_2, \ldots$ be independent random variables and $\nu$ be such that $X_1, \ldots, X_\nu$ are identically distributed with distribution function $F_0$ and $X_{\nu+1}, \ldots$ are identically distributed with distribution function $F_1$. Then the point $\nu$ is the change-point.

Change-point problems, in general, are concerned with the detection of the change-points and the estimation of their locations (see Shaban (1980) and Basseville (1988) for a list of references). Until recently, most of the work has been done on the problem of observing a process sequentially and detecting a change as soon as it occurs. This problem requires us to devise a stopping rule that rarely raises a false alarm and stops soon after the change occurs. Some of the procedures that have been studied for this problem are the CUSUM test (Page (1954), Lorden (1971) and the Shiryayev-Roberts (Shiryayev (1963), Roberts (1966)) procedure. These requirements for the stopping rule can be stated in terms of their expected run lengths. Under appropriate definitions of optimality, the CUSUM test and the Shiryayev-Roberts procedure have been proved to be optimal (cf. Pollak (1985,1987)).
Lately, fixed sample change-point problems, in which the process is observed over a fixed interval, have also been studied. In the fixed sample case we are interested in the problem of detecting any change-point that exists and also estimating the location of the change-points. Likelihood ratio and non-parametric methods have been used for fixed sample problems (cf. Sen and Srivastava (1975), Lombard (1986)). In this thesis we will be dealing with a fixed sample problem. We want to derive consistency results for multiple change-point problems.

1.1 Statement of the Problem

A multiple change-point problem deals with processes having more than one change-point. One problem of interest is the consistent estimation of the number and the locations of the change-points. The problem is described more precisely below.

Let \( \{X_i^{(n)}; 1 \leq i \leq n < \infty\} \) be an array of independent random variables (vectors) and let \( F_i^{(n)} \) be the distribution function of the random variable (vector) \( X_i^{(n)} \). Let the \( n^{th} \) row of the array have \( m_n \) change-points, that is, there exist \( \nu_1^{(n)}, \ldots, \nu_{m_n}^{(n)} \), where \( 0 = \nu_0^{(n)} < \nu_1^{(n)} < \ldots < \nu_{m_n}^{(n)} < \nu_{m_n+1}^{(n)} = n \), and distribution functions \( G_j \), where \( G_j \neq G_{j+1} \), for all \( j \geq 0 \), such that

\[
F_i^{(n)} = G_j, \text{ if } \nu_j^{(n)} < i \leq \nu_{j+1}^{(n)}, \ j = 0, \ldots, m_n. \tag{1.1}
\]

We are interested in the problem of consistent estimation of the number (\( m_n \)) of change-points and their locations \( \{\nu_j^{(n)}, 1 \leq j \leq m_n\} \). Suppose the class of distribution functions \( \{G_j\} \) belongs to a parametric family, then the estimation of the parameters is also of interest.

In order to study the structure of the problem we will consider the case of a univariate Gaussian process in detail, that is, the random variable \( X_i^{(n)} \) has a normal distribution with mean \( \mu_i^{(n)} \) and standard deviation \( \sigma_i^{(n)} \). In particular we will study the following problem.
Let \( \{\theta_j; 0 \leq j < \infty\} \) and \( \tau \) such that \( \theta_j \neq \theta_{j+1}, \forall j \) and

\[
\mu^{(n)}_i = \theta_j, \text{ if } \nu_j^{(n)} < i \leq \nu_{j+1}^{(n)}, \quad j = 0, \ldots, m_n
\]

\[
\sigma^{(n)}_i = \tau, \text{ for all } i \text{ and } n. \tag{1.2}
\]

Then the problem reduces to finding the number and the locations of the changes in the mean function of the process when the common variance may or may not be known. The estimation of the mean function is also of interest.

### 1.2 Review

Examples of change-point problems occur in quality control and epidemiology. Several authors have considered multiple change-points problems. Worsley and Srivastava (1986) use the likelihood ratio method to estimate changes in the mean of multivariate normal random variables. They apply their procedure to the mining data of Chernoff (1973) to estimate five change-points. Churchill (1989) studies a multiple change-point problem in genetics where a change-point denotes a change in the frequency of occurrence of certain base pairs. Some non-parametric methods have also been studied for estimating the change-points under location shift alternatives (Lombard (1986), Pollak and Gordon (unpublished manuscripts)).

The problem of detecting the changes in normal mean has been studied by several authors (Sen and Srivastava (1975), Hawkins (1977), Worsley (1979)). The literature on the asymptotic theory for this problem deals mainly with the case of a fixed number of changes in mean, say \( m \). Vostrikova (1981) and Yao (1988) consider two different estimation procedures to estimate the number and the locations of the change-points. They obtain certain consistency results for their estimates under the condition that \( \nu^{(m)}_i/n \) converges to \( p_i \), for all \( 1 \leq i \leq m \), where \( 0 < p_1 < \ldots < p_m < 1 \) are fixed constants. We would like to weaken the above conditions on the change-points for
these two procedures. We also are interested in using sequential methods for this problem and derive appropriate consistency results.

1.3 Outline

We will be discussing two different classes of estimation procedures. One is the class of procedures based on the ‘entire’ sample and the other is the class of pseudo-sequential procedures. Both the classes deal with a fixed sample only.

A procedure from the class of procedures based on the ‘entire’ sample, uses fixed sample tests (when there is a test involved) to detect change-points, where as, the one from the class of pseudo-sequential procedures use sequential tests. The main feature of the first class of procedures is that the estimates are invariant under the reversal of the order of occurrence of the observations, that is, they give the same results regardless of whether the sample is ordered 1, . . . , n or n, . . . , 1. Since the pseudo-sequential procedures use sequential tests, the sample is treated as ordered. So, the estimates in the pseudo-sequential case depend on the order in which the sample was used. Also, while the estimate of the location of any change-point depends on the entire sample in the former case, it depends only on the data before the estimated value in the latter. Since the distinguishing feature of the procedures in first class is the estimates being a function of the entire sample, we named them accordingly. The latter are named so for their use of sequential tests. Since they are not truly sequential, that is, we are dealing with a fixed sample that has already been observed, we added the term pseudo.

We will study two procedures that belong to the class of procedures based on the ‘entire’ sample. These are two procedures that have been used before. One is the binary segmentation procedure (Sen and Srivastava (1975)) and the other is the method by Yao (1988) using Schwarz’ Bayesian information criterion (Schwarz
(1978)). We will discuss these two procedures in Chapters 2 and 3 respectively.

In Section 2.1, we will describe the binary segmentation procedure. Vostrikova (1981), for a fixed number of change-points, proved the consistency of this procedure. We will prove a consistency theorem for the procedure in Section 2.2, under weaker conditions on the number and the locations of the change-points. In Section 2.3 we will give an example where this procedure fails asymptotically. This example gives us a necessary set of conditions for the procedure to be consistent, which we shall see is much stronger than the sufficient conditions of other procedures.

Chapter 3 deals with the other procedure that belongs to the first class. It was introduced by Yao (1988) and uses Schwarz' Bayesian information criterion to detect the change-points and estimate their locations. Yao proves the consistency of the procedure in the case of fixed number of change-points. In Section 3.1 we will describe the procedure and in Section 3.2 we will prove a consistency result for this procedure under weaker conditions on the number and the locations of the change-points. But the procedure is computationally infeasible even for moderate data size and more than four or five change-points.

In Chapter 4 we will introduce the pseudo-sequential procedures for the general problem we described in (1.1). In Section 4.1 we will describe the structure of this class of procedures. We will give a prototype of the consistency results for this class of procedures and the general structure of proof in Section 4.2. The description as well as the consistency statement can be used for the problem described in (1.2). The statistics used in the sequential tests (stopping rules) can be derived either by Bayesian arguments or by maximum likelihood arguments. These give us the Bayesian and the likelihood ratio methods, which we will discuss in the next two chapters.

We will deal with the Bayesian method in Chapter 5. These procedures are motivated by the Shiryayev-Roberts scheme for detecting a change-point. We will use the method developed by Pollak and Siegmund (1991) to detect a change in the
CHAPTER 1. INTRODUCTION

mean of a normal distribution, from an unknown initial mean. We will derive the statistic used for the known variance case in Section 5.1. and prove a consistency result for this procedure in Section 5.2. We will deal with the unknown variance case in Section 5.3, where we derive the test statistic and describe the difficulty we face in the implementation of the procedure. We will also sketch the proof a consistency theorem.

In Chapter 6 we will consider the likelihood ratio method. These are the procedures motivated by the CUSUM type tests for detecting a change-point. In this chapter we will consider, in detail, both the known and the unknown variance case. In Section 6.1 we will describe the test statistics used for this procedure and derive consistency results. We will discuss the unknown variance case in Section 6.2 and derive appropriate consistency results.

Chapter 7 has Monte Carlo results comparing the small sample performance of the procedures. In Section 7.1 we will discuss the advantages and disadvantages of the various procedures through numerical study. We will describe some ad hoc modifications to some of the procedures in order to improve their efficiency. We will discuss the improvements in their performance. We will devise a factorial experiment to analyze the effects of various factors that contribute to the detection of the change-points and report the results in Section 7.2. We will use some tail probability approximations to determine the critical levels for various test statistics we use in the different procedures. In Section 7.3 we will use all the procedures we discussed to estimate the number and the locations of change-points in the Nile data (from Cobb (1978)). The data gives the annual volume of the Nile from the years 1871 to 1970.

Chapter 8 deals with a related problem. We will develop the Bayesian procedure for detecting changes in the rate of an exponential process and prove the consistency of the procedure. We will then apply the procedure to British coal-mining disaster data (from Jarrett (1979)). It gives the time interval in days between consecutive
exploded in mines, from 15 March, 1851 to 22 March, 1962. In all there were 191 explosions involving 10 or more men killed.

For the implementation of the test procedures we needed some tail probability approximations to derive the critical levels of the test statistics. In Section A.1 we will discuss the tail probability approximations used in obtaining critical values and in Section A.2 we will derive two of the approximations by maxima of random fields arguments (cf. Siegmund 1988).

From the developments in the following chapters we will see that the pseudo-sequential procedures perform considerably better than the methods based on the 'entire' sample. The problem of overestimating the locations of the change-points can be easily rectified by using the likelihood function of the data between consecutive stopping times or by using the posterior distribution of the location of the change-points. The Bayesian scheme, where it is possible to compute one, performs better than the likelihood ratio method, since it doesn’t let the outliers affect the data. The likelihood ratio method can be easily described for any class of distributions. We can conclude that the pseudo-sequential procedures have better performance for our estimation problem.
Chapter 2

Binary Segmentation

In the outline we described the two class of procedures we will be studying. Since the procedures belonging to the class of methods based on the ‘entire’ sample have been studied before we will discuss it first. In this chapter we will consider the first of the two procedures that belong to this class. The binary segmentation procedure appears to have been introduced by Sen and Srivastava (1975) to detect a change in the mean of normal random variables. We will now describe the procedure.

2.1 Description of the Procedure

Let $Y_1, \ldots, Y_d$ be the data. The binary segmentation procedure works as follows. We will test the hypothesis that $Y_1, \ldots, Y_d$ come from the same distribution against the alternative that there is a $\nu$ such that $Y_1, \ldots, Y_\nu$ and $Y_{\nu+1}, \ldots, Y_d$ have different distributions. If we accept the hypothesis that there is such a $\nu$, we call it a change-point and estimate it through maximum likelihood. We then split the data into two segments at the estimated change-point and continue the procedure on the two subsamples (hence the name binary segmentation). We continue like this until we no longer detect any change-point in any of the segments. We then estimate the number
of change-points and their locations by the number of change-points detected and the maximum likelihood estimates of their locations.

We will now study the binary segmentation procedure for the Gaussian case, where the array of random variables satisfy equation (1.2). We will assume that the common variance $\tau^2$ is known, and hence without loss of generality we can take $\tau^2$ to be 1.

Let $Y_1, \ldots, Y_d$ be independent normal random variables with unit variance. The problem of testing the hypothesis that $Y_1, \ldots, Y_d$ have the same distribution against the alternative that $Y_1, \ldots, Y_i$ and $Y_{i+1}, \ldots, Y_d$ have different distributions reduces to the problem of testing for equality of the means of the two groups. We will use the two sample generalized likelihood ratio statistic for a given value of $i$

$$Z_i = (iS_d/d - S_i)/\sqrt{i(1 - i/d)},$$

(2.1)

where $S_k = Y_1 + \cdots + Y_k$. Now to test for the hypothesis that $Y_1, \ldots, Y_d$ have the same distribution against the alternative that there exists a $\nu$ such that $Y_1, \ldots, Y_{\nu}$ and $Y_{\nu+1}, \ldots, Y_d$ have different means we use the statistic:

$$S(Y_1, \ldots, Y_d) = \max_{0 < i < d} |Z_i|.$$  

(2.2)

If $S(Y_1, \ldots, Y_d)$ is larger than the critical level $b_d$, which depends on $d$, we reject the hypothesis of no change-point. We estimate the change-point to be any $k$ such that $|Z_k| = S(Y_1, \ldots, Y_d)$. This divides the data into two segments $Y_1, \ldots, Y_k$ and $Y_{k+1}, \ldots, Y_d$. We will repeat the procedure on the two segments and all the subsequent segments. The critical levels for the segments can either be the same as the one chosen for the whole sample or can be modified to reflect the size of the subsample. We will stop when we no longer detect a change-point in any of the segments. Suppose at the end of this we have $l$ segments (clearly $l$ is at least one and at most $d$). Then the estimated number of change-points is $l - 1$. The locations of these change-points are estimated by the right end point of the segments (except $d$). We will now derive a consistency result for this procedure.
2.2 Consistency Result

We are interested in the consistency of the binary segmentation procedure. The consistency of this procedure when there are a fixed number of change-points, say $r$, satisfying the condition $\nu_i^{(n)}/n$ converges to $p_i$ where $0 < p_1 < \ldots < p_r < 1$ follows from Vostrikova (1981). In Theorem 2.1 we will derive the consistency of this procedure under weaker conditions on the number and the locations of the change-points.

Let $n$ be fixed. We will drop the superscript $(n)$ for notational simplicity. Let $S_k = X_1 + \cdots + X_k$, $1 \leq k \leq n$. Then for $0 \leq k_1 < k < k_2 \leq n$ the generalized likelihood ratio statistic for any $k$ in the segment $\{k_1 + 1, \ldots, k_2\}$ is given by

$$Z_{k_1,k_2}^k = \left[ \frac{k - k_1}{k_2 - k_1} (S_{k_2} - S_{k_1}) - (S_k - S_{k_1}) \right] / \sqrt{(k - k_1) \left( 1 - \frac{k - k_1}{k_2 - k_1} \right)} \quad (2.3)$$

Let $U_i = X_i - \mu_i, 1 \leq i \leq m$ and $V_k = U_1 + \cdots + U_k$. Let $M_k = \mu_1 + \cdots + \mu_k$. Then observe that $Z_{k_1,k_2}^k = W_{k_1,k_2}^k + \Theta_{k_1,k_2}^k$, where

$$W_{k_1,k_2}^k = \left[ \frac{k - k_1}{k_2 - k_1} (V_{k_2} - V_{k_1}) - (V_k - V_{k_1}) \right] / \sqrt{(k - k_1) \left( 1 - \frac{k - k_1}{k_2 - k_1} \right)} \quad (2.4)$$

and

$$\Theta_{k_1,k_2}^k = \left[ \frac{k - k_1}{k_2 - k_1} (M_{k_2} - M_{k_1}) - (M_k - M_{k_1}) \right] / \sqrt{(k - k_1) \left( 1 - \frac{k - k_1}{k_2 - k_1} \right)} \quad (2.5)$$

Let $\Theta_{k_1,k_2}$ be defined as

$$\Theta_{k_1,k_2} = \max_{k_1 \leq k \leq k_2} \left| \Theta_{k_1,k_2}^k \right| \quad (2.6)$$

We have the following consistency theorem for the binary segmentation procedure.

**Theorem 2.1** Let $b_n = n^{3/8}$ be the boundary. Let $\hat{m}_n$ and $\hat{\nu}_1^{(n)}, \ldots, \hat{\nu}_{\hat{m}_n}^{(n)}$ respectively be the estimated number and locations of the change-points. Suppose the following
CHAPTER 2. BINARY SEGMENTATION

conditions hold:

(i) \( \inf_{i} \{ \nu_{i+1}^{(n)} - \nu_{i}^{(n)} \} \geq 2cn^{1-\beta} \forall n, \) where \( c > 0 \) and \( 0 \leq \beta < \frac{1}{8} \) are independent of \( n, \)

(ii) \( \inf_{i} |\theta_{i+1} - \theta_{i}| \geq \delta \) for some \( \delta > 0, \)

(iii) \( \sup_{i} |\theta_{i}| \leq B \) for some \( B > 0. \)

Then \( P(A_n) \) converges to 1, where

\[ A_n = \{ \hat{m}_n = m_n; \ |\hat{\nu}_{j}^{(n)} - \nu_{j}^{(n)}| \leq n^{3/4}, 1 \leq j \leq m_n \}. \]

Remark 2.1 Condition (i), which is a condition on the minimal distance between change-points, is less severe for other procedures that we will be discussing, than in this theorem.

Remark 2.2 A counter example in Section 2.3 shows the necessity of a larger distance between change-points for this procedure than the other procedures.

The result is proved using a series of lemmas which follow.

Lemma 2.1 Let \( W \) be as defined by (2.4). Then \( P(B_n) \rightarrow 1 \) as \( n \rightarrow \infty, \) where

\[ B_n = \left\{ \max_{0 \leq k_1 < k < k_2 \leq n} |W_{k_1,k_2}^k| \leq 3\sqrt{\log n} \right\}. \]

Proof: Observe that for all \( 0 \leq k_1 < k < k_2 \leq n, W_{k_1,k_2}^k \) is standard normal. So

\[
P \left( \max_{0 \leq k_1 < k < k_2 \leq n} |W_{k_1,k_2}^k| > 3\sqrt{\log n} \right) \leq \sum_{0 \leq k_1 < k < k_2 \leq n} P \left( |W_{k_1,k_2}^k| > 3\sqrt{\log n} \right)
\]

\[
\leq n^3 \Phi \left( -3\sqrt{\log n} \right) \rightarrow 0 \text{ as } n \rightarrow \infty.
\]

Since \( P(B_n^c) \rightarrow 0 \) as \( n \rightarrow \infty \) the lemma follows. \( \square \)
Lemma 2.2 Let \( l > 0 \) be an integer. Let \( 0 = a_0 < a_1 < \ldots < a_l < a_{l+1} = 1 \) and let \( \lambda_0, \ldots, \lambda_l \) be such that, for all \( i, \lambda_i \neq \lambda_{i+1} \) and \( \sum_{i=0}^{l} (a_{i+1} - a_i) \lambda_i = 0 \). Let the function \( f \) be defined by
\[
f(x) = \frac{\sum_{j=1}^{l} (a_j - a_{j-1}) \lambda_j - 1 + (x - a_i) \lambda_i}{\sqrt{x(1-x)}} \quad \text{for } a_i \leq x \leq a_{i+1}, \quad 0 \leq i \leq l.
\] (2.7)

Let \( f^* = \max_{0 \leq x \leq 1} |f(x)| \). Let \( x \) be such that \( |f(x)| = f^* \). Then there exists \( i \) such that \( 1 \leq i \leq l \) and \( a_i = x \), that is, the maximum of \( |f| \) can be attained at one of the \( a_i \)'s only.

Proof: Observe that \( f \) is non-zero for some \( x \) and hence \( f^* \) is positive. Since \( f \) and \( -f \) have similar structure, we can assume without loss of generality that \( f^* = \max_{0 \leq x \leq 1} f \). Now observe that \( f \) is either strictly monotonic or identically zero in the range \([0, a_1]\) and \([a_l, 1]\). Since \( f(0) = f(1) = 0 \), in both the ranges if \( f^* \) is attained then it has to be at \( a_1 \) and \( a_l \) respectively. So it is enough to prove the following.

Claim: Let \( g \) be a function of the form \( g(x) = \frac{ax + b}{[x(1-x)]^{1/2}} \) over the range \([c, d]\), where \( 0 < c < d < 1 \). If \( g(x) \) is positive for some \( x \in [c, d] \), then for all \( c < x < d \), \( g(x) < \max\{g(c), g(d)\} \).

Since multiplying by a positive constant does not affect the maxima, we could assume without loss of generality that \( a = 0 \), 1 or \(-1\). Also, since \( g(x) \) is positive for some \( c \leq x \leq d \), and \( x(1-x) \) is positive for all \( x \) in \((0,1)\), we get that \( ax + b > 0 \) for some \( c \leq x \leq d \). Thus we have to prove the claim for the following three cases: (1) \( a = 0 \), \( b > 0 \), because if \( a = 0 \) then \( ax + b = b > 0 \) for some \( x \in [c, d] \) only if \( b > 0 \), (2) \( a = 1 \), \( b > -d \), because in this case \( ax + b = x + b > 0 \) for some \( x \in [c, d] \) only if \( b > -d \), and (3) \( a = -1 \), \( b > c \), since \( ax + b = b - x > 0 \) for some \( x \in [c, d] \), only if \( b > c \).

Case 1: This case is very simple. Observe that \( x(1-x) \) increases for \( x \in [0, 1/2] \) and
decreases for \( x \in [1/2, 1] \). So the function \( g \) is either strictly monotonic or decreases initially and then increases and in both the cases the claim follows.

**Case 2:** In this case the derivative of \( g \) is 
\[
\frac{x(1+2b) - b}{2[x(1-x)]^{3/2}}.
\]
Let \( x_0 = b/(1+2b) \). Observe that (i) if \( 1+2b > 0 \), then \( g' \) is negative for \( \{x < x_0\} \cap (0,1) \) and positive for \( \{x > x_0\} \cap (0,1) \) and (ii) if \( 1+2b < 0 \) then, \( g' \) is negative for \( \{x > x_0\} \cap (0,1) \) and positive for \( \{x < x_0\} \cap (0,1) \). So if (i) holds then \( g \) decreases until \( x_0 \) and then increases, in which case the maximum is attained at one of the end points. This takes care of the case \( b \geq -1/2 \). So we are left with the case \( -d < b < -1/2 \). Since \( x_0 > 1 \) for this range of \( b \), by (ii) we get that \( g \) increases in the range \( c \leq x \leq d \) and hence the claim is true.

**Case 3:** Here the derivative of \( g \) is 
\[
\frac{-x(1+2b) - b}{2[x(1-x)]^{3/2}}.
\]
Observe that, for all \( b > 0 \), the derivative \( g' \) is negative for all \( x > 0 \). Hence we get that \( g \) decreases in the range \( c \leq x \leq d \) and the claim follows. 

We have the following setup that will be used in Lemmas 2.3 through 2.8. Let \( k_1, k_2 \) be such that
\[
\nu_{i_0} \leq k_1 < \nu_{i_0+1} < \ldots < \nu_{i_0+l} < k_2 \leq \nu_{i_0+l+1},
\]  
(2.8)
where \( 0 \leq i_0 \leq m_n - l \). Since the function \( \Theta \) as defined by (2.5) is invariant under location shift we may assume without loss of generality that
\[
M_{k_2} - M_{k_1} = 0.
\]  
(2.9)
Since a location shift doesn't affect the change-points, in the following lemmas when we refer to the mean of a random variable \( X_k \) within the context of segment \( \{k_1 + 1, \ldots, k_2\} \), we deal not with the actual mean but the mean after a location shift so that equation (2.9) holds.

**Lemma 2.3** Let \( k_1, k_2 \) be such that equation (2.8) holds for some \( l > 0 \). Then if \( \left| \Theta_{k_1,k_2}^{+} \right| = \Theta_{k_1,k_2} \), then \( k = \nu_{i_0+i} \) for some \( 0 \leq i \leq l \).
CHAPTER 2. BINARY SEGMENTATION

**Proof:** Let \( a_i = (\nu_{i_0+i} - k_1)/(k_2 - k_1) \) and \( \theta_i = EX_{\nu_{i_0+i}}. \) Then
\[
\Theta_{k_1,k_2}^k = f\left( \frac{k - k_1}{k_2 - k_1} \right) \sqrt{k_2 - k_1}.
\]
Now the result follows from Lemma 2.2. \( \Box \)

We need one or both of the following conditions on \( k_1, k_2 \) in Lemmas 2.4 through 2.7.
\[
k_1 < \nu_{i_0+i} - cn^{1-\beta} < \nu_{i_0+i} + cn^{1-\beta} < k_2 \text{ for some } 1 \leq i \leq l. \tag{2.10}
\]
\[
[(\nu_{i_0+i-1} - k_1) \wedge (k_1 - \nu_{i_0})] \vee [(\nu_{i_0+i+1} - k_2) \wedge (k_2 - \nu_{i_0+i})] \leq n^\alpha \log n. \tag{2.11}
\]

**Lemma 2.4** Suppose \( k_1, k_2 \) satisfy (2.10) and the conditions of the theorem hold. Then
\[
\max_{k_1 < k < k_2} \left| \Theta_{k_1,k_2}^k \right| \geq \frac{\delta}{2} cn^{1/2-\beta}.
\]

**Proof:** Let \( EX_{\nu_1} = \theta \) and \( EX_{\nu_{i+1}} = \theta' \). Since \( |\theta - \theta'| > \delta \), by condition (ii) of the theorem, we get that
\[
|\theta| \vee |\theta'| > \delta/2. \tag{2.12}
\]
By (2.12) and the definition of \( M \) we can see that
\[
|M_{\nu_i} - M_{\nu_i-cn^{1-\beta}}| \vee |M_{\nu_i+cn^{1-\beta}} - M_{\nu_i}| > \frac{\delta}{2} cn^{1-\beta}. \tag{2.13}
\]
From (2.13) we get that the maximum of \( |M_k - M_{k_1}| \) is at least \( \frac{\delta}{4} cn^{1-\beta} \). Finally since \( (k - k_1)(k_2 - k)/(k_2 - k_1) \leq (k_2 - k_1)/4 \leq n/4 \) we get the lemma. \( \Box \)

**Lemma 2.5** Suppose \( k_1, k_2 \) satisfy (2.10) and (2.11) for some \( \alpha < 1 - 2\beta \). Let \( \nu \) be a change-point that satisfies
\[
\Theta_{k_1,k_2}^\nu > \Theta_{k_1,k_2} - 6\sqrt{\log n}. \tag{2.14}
\]
Let \( \Theta_{k_1,k_2} = a/\sqrt{(\nu - k_1)(k_2 - \nu)/(k_2 - k_1)} \). Then under the conditions of the theorem, for large \( n \),
\[
a > \frac{\delta cn^{1-\beta}}{5}. \tag{2.15}
\]
CHAPTER 2. BINARY SEGMENTATION

\textbf{Proof:} Observe that
\[ \Theta'_{k_1,k_2} = a/\sqrt{(\nu - k_1)(k_2 - \nu)/(k_2 - k_1)} \leq 2\sqrt{2}B\sqrt{(\nu - k_1) \wedge (k_2 - \nu)}. \]

Also \((\nu - k_1) \wedge (k_2 - \nu)\) is either less than \(n^\alpha \log n\) or larger than \(2cn^{1-\beta} - n^\alpha \log n\). Since by Lemma 2.4 the first possibility contradicts (2.14) (because \(\alpha < 1 - 2\beta\)), for large \(n\), we get that
\[ (\nu - k_1) \wedge (k_2 - \nu) \geq 2cn^{1-\beta} - n^\alpha \log n. \] (2.16)

Without loss of generality we could assume that \(2(\nu - k_1) \leq (k_2 - k_1)\). Let \(\nu = \nu_{i_0+i}\) for some \(1 \leq i \leq l\). Let \(\theta\) and \(\theta'\) be the mean of \(X_\nu\) and \(X_{\nu+1}\) respectively. Then by condition (ii) of the theorem either \(|\theta|\) or \(|\theta'|\) is greater than \(\delta/2\).

Suppose \(|\theta| > \delta/2\) and (2.15) does not hold. We will consider two cases.

\textbf{Case} \(i = 1\). Since \(a = \theta(\nu - k_1)\), by (2.16) we get that \(a > \delta cn^{1-\beta}/2\) for large \(n\).

\textbf{Case} \(i > 1\): Let \(\nu' = \nu_{i_0+i-1}\). If \(a < \delta cn^{1-\beta}/3\), then
\[ |\Theta'_{k_1,k_2}| = \frac{|a - \theta(\nu' - \nu)|}{\sqrt{(\nu' - k_1)(k_2 - \nu')/(k_2 - k_1)}} \geq \frac{2\delta cn^{1-\beta}/3}{\sqrt{(\nu - k_1)(k_2 - \nu)/(k_2 - k_1)}} \geq 2|\Theta_{k_1,k_2}'|.
\]

For large \(n\), by Lemma 2.4, this contradicts (2.14). Hence \(a > \delta cn^{1-\beta}/3\) for large \(n\).

Now suppose \(|\theta'| > \delta/2\) and (2.15) is not true. We will again consider two cases.

\textbf{Case} \(i = l\): Since \(a = |\theta'|(k_2 - \nu)\), by (2.16) we get that \(a > \delta cn^{1-\beta}/2\) for large \(n\).

\textbf{Case} \(i < l\). Let \(\nu' = \nu + cn^{1-\beta}\). If \(a < \delta cn^{1-\beta}/5\), then
\[ |\Theta_{k_1,k_2}'| = \frac{|a + \theta(\nu' - \nu)|}{\sqrt{(\nu' - k_1)(k_2 - \nu')/(k_2 - k_1)}} \geq \frac{3\delta cn^{1-\beta}/10}{\sqrt{(\nu' - k_1)(k_2 - \nu)/(k_2 - k_1)}} \geq \frac{3}{2\sqrt{2}} |\Theta_{k_1,k_2}'|. \]
Again, by Lemma 2.4, for large \( n \), this contradicts (2.14). Thus \( a \) is larger than \( \delta cn^{1-\beta}/5 \) for large \( n \). \[ \square \]

**Lemma 2.6** Suppose \( k_1, k_2 \) satisfy (2.10) and (2.11) for some \( \alpha < 1 - 2\beta \). Let \( \nu \) be a change-point that satisfies (2.14). Then under the conditions of the theorem, for some \( 0 < l < n^\alpha \log n \)

\[
\Theta_{k_1,k_2}^\nu > \Theta_{k_1,k_2}^{\nu+l} + 6 \sqrt{\log n}.
\]

**Proof:** By (2.16) both \( \nu - k_1 \) and \( k_2 - \nu \) are larger than \( 2cn^{1-\beta} - n^\alpha \log n \). Let \( \nu' > \nu \) be the next change-point (\( \nu' = k_2 \) if there isn't any). We will consider two cases (1) \( \nu' = k_2 \) and (2) \( \nu' < k_2 \).

**Case 1:** Let \( i = \nu - k_1 \) and \( h = k_2 - \nu \).

Since \( \Theta_{k_1,k_2}^\nu = \frac{a \sqrt{i + h}}{\sqrt{i} \sqrt{h}} \) we get that \( \Theta_{k_1,k_2}^{\nu+l} = \frac{(h - l)}{h} \frac{a \sqrt{i + h}}{\sqrt{(i + l)(h - l)}} \).

So, if \( E_l = \Theta_{k_1,k_2}^\nu - \Theta_{k_1,k_2}^{\nu+l} \), we get

\[
E_l = \frac{a \sqrt{i + h}}{h} \left( \sqrt{\frac{h}{i}} - \sqrt{\frac{h - l}{i + l}} \right)
\]

\[
= \frac{a \sqrt{i + h}}{h} \frac{h - l}{i + l} \left( \sqrt{\frac{h}{i}} + \sqrt{\frac{h - l}{i + l}} \right)
\]

\[
= \frac{la(i + h)^{3/2}}{h \sqrt{i} \sqrt{i + l} (\sqrt{(i + l)h + i(h - l)})}
\]

The last inequality follows from the fact that \( n^\alpha \log n < i \) for large \( n \). Since

\[
E_l \geq \frac{\kappa la}{(i \wedge h)^{3/2}}, \quad a > \frac{\delta cn^{1-\beta}}{5} \quad \text{and} \quad i \wedge h \leq n,
\]

we get that \( E_{n^\alpha \log n} \geq \kappa' n^{\alpha + 1 - \beta - 3/2} \geq \kappa' \log n \) if \( \alpha \geq 1/2 + \beta \). (\( \kappa \) and \( \kappa' \) are constants independent of \( n \)).
Case 2: Let $i = \nu - k_1$, $h = cn^{1-\beta}$ and $j = k_2 - \nu - h$. Since there is change-point between $\nu$ and $k_2$, $j$ is at least as large as $h$. By (2.16), for large $n$, $h$ is smaller than $i$.

Let $\Theta^\nu_{k_1,k_2} = \frac{a\sqrt{i+j+h}}{\sqrt{i(j+h)}}$, $\Theta^{\nu+h}_{k_1,k_2} = \frac{(a+h\theta)\sqrt{i+j+h}}{\sqrt{(i+j)h}}$ and $b = \Theta^{\nu+h}_{k_1,k_2} - \Theta^\nu_{k_1,k_2}$.

Then $\theta$ is the solution of $\Theta^{\nu+h}_{k_1,k_2} = \Theta^\nu_{k_1,k_2} + b$, that is,

$$\theta = \frac{a\sqrt{(i+h)j}}{h} \left[ \frac{1}{\sqrt{i(j+h)}} - \frac{1}{\sqrt{(i+h)j}} + \frac{b}{a\sqrt{i+j+h}} \right].$$

Let $E_l = \Theta^\nu_{k_1,k_2} - \Theta^{\nu+l}_{k_1,k_2}$. Then for $0 < l < h$,

$$E_l \geq \frac{a\sqrt{i+j+h}}{\sqrt{i(j+h)}} - \frac{(a+l\theta)\sqrt{i+j+h}}{\sqrt{(i+l)(j+h-l)}}$$

$$= \frac{a\sqrt{i+j+h}}{\sqrt{i+l}(j+h-l)} \left[ \frac{l\sqrt{(i+h)j}}{h} \left\{ \frac{1}{\sqrt{i(j+h)}} - \frac{1}{\sqrt{(i+h)j}} + \frac{b}{a\sqrt{i+h+j}} \right\} \right]$$

$$= \frac{a\sqrt{i+j+h}}{\sqrt{i+l}(j+h-l)} \left[ \frac{(\sqrt{(i+l)(j+h-l)} - \sqrt{i(j+h)}}{\sqrt{i(j+h)}} \right]$$

$$- \frac{a\sqrt{i+j+h}}{\sqrt{i+l}(j+h-l)} \frac{l}{h} \left[ \frac{(\sqrt{(i+h)j} - \sqrt{i(j+h)}}{\sqrt{i(j+h)}} \right]$$

$$- \frac{a\sqrt{i+j+h}}{\sqrt{i+l}(j+h-l)} \frac{bl\sqrt{(i+h)j}}{ha\sqrt{i+j+h}}$$

$$= \frac{(h-l)}{h} \frac{a\sqrt{i+j+h}}{\sqrt{(i+l)(j+h-l)}} \left( \frac{(i+l)(j+h-l) - \sqrt{i(j+h)}}{\sqrt{i(j+h)}} \right)$$

$$+ \frac{l}{h} \frac{a\sqrt{i+j+h}}{\sqrt{(i+l)(j+h-l)}} \frac{\sqrt{(i+l)(j+h+l)} - \sqrt{(i+h)j}}{\sqrt{i(j+h)}}$$

$$- \frac{bl}{h} \frac{\sqrt{(i+h)j}}{\sqrt{(i+l)(j+h-l)}}$$

(2.17) (2.18) (2.19)
Observe that (2.17) and (2.18) reduce to
\[
\begin{align*}
\frac{(h - l)}{h} & \frac{a \sqrt{i + j + h}}{\sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)} \sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}}} \frac{l(j + h - i - l)}{\sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)} (\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}})} \\
& = \frac{a \sqrt{i + j + h}}{\sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)} (\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}})} \\
& + \frac{(h - l)}{h} \frac{a \sqrt{i + j + h}}{\sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)} (\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}})}
\end{align*}
\]
(2.20)
and
\[
\begin{align*}
\frac{l}{h} & \frac{a \sqrt{i + j + h}}{\sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)} (\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}})} \frac{(h - l)(i + l - j)}{\sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)} (\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}})}
\end{align*}
\]
(2.21)
respectively. Now adding (2.21) and (2.22) would give us
\[
\frac{a \sqrt{i + j + h}}{h \sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)}} \sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}}
\]
\[
\times \left[ \frac{1}{(\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}})} - \frac{1}{(\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}})} \right]
\]
\[
\times \frac{1}{\sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)} (\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}}) \sqrt{i(j + h)} + \sqrt{(i + h)j}}
\]
(2.23)
Finally the sum of (2.20) and (2.23) reduces to $E_{1l} \times (1 + E_{2l})$, where
\[
E_{1l} = \frac{a \sqrt{i + j + h}}{\sqrt{i(j + h)} \sqrt{(i + l)(j + h - l)} (\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}})}
\]
\[
E_{2l} = \frac{(j - i)(j - i - l)}{(\sqrt{(i + l)(j + h - l) + \sqrt{i(j + h)}}) (\sqrt{i(j + h)} + \sqrt{(i + h)j})}
\]
Thus $E_{l} = E_{1l} \times (1 + E_{2l}) + E_{3l}$, where
\[
E_{3l} = -\frac{bl}{h} \frac{\sqrt{(i + h)j}}{\sqrt{(i + l)(j + h - l)}}.
\]
Observe that \( l \) is less than \( h/2 \) for large \( n \). So
\[
E_{1l} \geq \frac{al(h-l)\sqrt{1 \lor (j+h)}}{\sqrt{i(j+h)}\sqrt{2i(j+h)} \left( \sqrt{2i(j+h)} + \sqrt{i(j+h)} \right)} - \frac{\kappa al(h-l)}{[i \lor (j+h)]^{3/2}[i \land (j+h)]}.
\]
Observe also that \( a > \delta cn^{1-\beta}/5 \) by Lemma 2.5 and \( i \lor (j+h) < n \). Hence we get that
\[
E_{1n^\alpha \log n} \geq \kappa' n^{\alpha+2-2\beta-5/2} \log n \geq \kappa' \log n
\]
provided \( \alpha \geq 1/2 + 2\beta \), where \( \kappa \) and \( \kappa' \) are non-negative and are independent of \( n \).

Since \( i, j, h \geq cn^{1-\beta} \) we get that
\[
E_{2l} \geq \frac{-l^2/2}{\eta n^{2(1-\beta)}} \geq -\eta' n^{2\alpha+2\beta-2} \log^2 n,
\]
where \( \eta \) and \( \eta' \) are non-negative and are independent of \( n \). So for large \( n \), \( E_{2l} \) is greater than \(-1/2 \) for all \( l \leq n^\alpha \log n \).

Finally, the largest value \( b \) can take is \( 6\sqrt{\log n} \). Since \( l \leq h \) for large \( n \),
\[
E_{3l} \geq -\omega 3n^{\alpha+\beta-1} \sqrt{\log n} \sqrt{2ij / \sqrt{ij}},
\]
where \( \omega > 0 \) is independent of \( n \). This lower bound converges to zero provided \( \alpha < 1 - \beta \).

Thus we get that \( E_{n^\alpha \log n} \) is larger than \( \zeta \log n \), for some positive constant \( \zeta \) independent of \( n \), for large \( n \), provided \( 1/2 + 2\beta < \alpha < 1 - 2\beta \). So if we set \( \alpha = 5/8 + \beta \) we get the lemma. \( \square \)

**Lemma 2.7** Let \((U_1, \ldots, U_n) \in B_n\). Let \( k_1, k_2 \) satisfy (2.10) and (2.11) for \( \alpha = 5/8 + \beta \). Let \( k_0 \) be such that
\[
Z_{k_1,k_2}^{k_0} = \max_{k_1 < k < k_2} |Z_{k_1,k_2}^k|.
\]
Then for some \( 1 \leq i \leq l \)
\[
|\nu_{i_0+i} - k_0| \leq n^{5/8+\beta} \log n.
\]
(2.24)
**Proof:** By Lemma 2.1 we get that \( \max_{k_1 < k < k_2} |Z_{k_1, k_2}^k| \leq 3\sqrt{\log n} \) on \( B_n \). So we get that

\[
Z_{k_1, k_2}^{k_0} \geq \Theta_{k_1, k_2} - 3\sqrt{\log n},
\]

and

\[
\Theta_{k_1, k_2}^{k_0} \geq Z_{k_1, k_2}^{k_0} - 3\sqrt{\log n}.
\]

Hence

\[
\Theta_{k_1, k_2}^{k_0} \geq \Theta_{k_1, k_2} - 6\sqrt{\log n}. \tag{2.25}
\]

So for large \( n \), \( \Theta_{k_1, k_2}^{k_0} \) is positive. Suppose (2.24) is not true. For some \( 1 \leq h \leq l - 1 \) let \( h \) be such that \( \nu_{i_0 + h} + n^\alpha \log n < k_0 < \nu_{i_0 + h + 1} - n^\alpha \log n \). By Lemma 2.2, for \( k \) in the range \( \{\nu_{i_0 + h}, \ldots, \nu_{i_0 + h + 1}\} \), \( \Theta_{k_1, k_2}^k \) either is monotonic or decreases and then increases and

\[
\Theta_{k_1, k_2}^\nu_{i_0 + h} \lor \Theta_{k_1, k_2}^\nu_{i_0 + h + 1} > \Theta_{k_1, k_2}^{k_0}.
\]

Suppose

\[
\Theta_{k_1, k_2}^{\nu_{i_0 + h}} > \Theta_{k_1, k_2}^{k_0}.
\]

Then by Lemma 2.6 we get that

\[
\Theta_{k_1, k_2}^k < \Theta_{k_1, k_2} - 6\sqrt{\log n},
\]

for some \( \nu_{i_0 + h} < k \leq \nu_{i_0 + h} + n^\alpha \log n \). Since \( k_0 > k \), we get that (because the mean function is in the increasing stage at \( k_0 \)),

\[
\Theta_{k_1, k_2}^\nu_{i_0 + h + 1} > \Theta_{k_1, k_2}^{k_0}.
\]

So again by Lemma 2.6

\[
\Theta_{k_1, k_2}^{k'} < \Theta_{k_1, k_2} - 6\sqrt{\log n},
\]

for some \( \nu_{i_0 + h + 1} > k' \geq \nu_{i_0 + h + 1} - n^\alpha \log n \). But since \( k' > k_0 \) we get a contradiction.

Hence (2.27) holds and the lemma follows. \( \square \)
Lemma 2.8 Suppose $k_1, k_2$ be such that one of the following conditions hold:

(i) $l = 1$ and $[\nu_{i_0 + 1} - k_1] \wedge [k_2 - \nu_{i_0 + 1}] \leq n^{5/8 + \beta} \log n$  

(2.26)

(ii) $l = 2$ and $[\nu_{i_0 + 1} - k_1] \lor [k_2 - \nu_{i_0 + 2}] \leq n^{5/8 + \beta} \log n$.  

(2.27)

Then for large $n$,

$$\Theta_{k_1, k_2} < n^{3/8} - 3\sqrt{\log n}.$$  

Proof: If condition (i) holds then by Lemma 2.3

$$\Theta_{k_1, k_2} = |\Theta_{k_1, k_2}^{\nu_{i_0 + 1}}| \leq 2\sqrt{2} B \sqrt{[\nu_{i_0 + 1} - k_1] \wedge [k_2 - \nu_{i_0 + 1}]} \leq 2\sqrt{2} B \sqrt{n^{5/8 + \beta} \log n}.$$  

If condition (ii) holds again by Lemma 2.3

$$\Theta_{k_1, k_2} = \left| \Theta_{k_1, k_2}^{\nu_{i_0 + 1}} \lor \Theta_{k_1, k_2}^{\nu_{i_0 + 2}} \right| \leq 2\sqrt{2} B \sqrt{[\nu_{i_0 + 1} - k_1] \lor [k_2 - \nu_{i_0 + 2}]} \leq 2\sqrt{2} B \sqrt{n^{5/8 + \beta} \log n}.$$  

Since $\beta < 1/8$ the lemma follows. \qed

From these lemmas we can prove the consistency theorem stated earlier.

Proof of the Theorem: We will get upper bounds for $P(A_n)$ and prove that the upper bounds converge to zero. Let $B_n$ be as defined in Lemma 2.1. Then we claim that for large $n$, $B_n \subseteq A_n$.

Since $k_1 = 0$ and $k_2 = n$ satisfy condition (2.11) (with any non-negative $\alpha$) by Lemma 2.7 every subsequent segment satisfies (2.11) with $\alpha = 5/8 + \beta$. If we have detected less than $m_n$ change-points then there exists a segment $\{k_1 + 1, \ldots, k_2\}$ such that (2.10) holds. So by Lemma 2.4, $\Theta_{k_1, k_2} > n^{3/8} + 3\sqrt{\log n}$ and hence a change-point will be detected in this segment. Thus $\hat{m}_n \geq m_n$. Once we detect $m_n$ change-points all the segments have end points that satisfy either (2.26) or (2.27). In either case by Lemma 2.8 we don't detect any change-point in the segment. Thus for
all \((U_1, \ldots, U_n) \in B_n\) we detect exactly \(m_n\) change-points and \(|\nu_i - \hat{\nu}_i| \leq n^{5/8+\beta} \log n\) for all \(1 \leq i \leq m_n\). Thus \(A_n \subset B_n\) and by Lemma 2.1 the theorem follows. Q.E.D.

The following results can be derived from the theorem.

**Corollary 2.1** Under the conditions of Theorem 2.1 the following are true.

\[(i) \quad P(m_n = m_n) \to 1 \text{ as } n \to \infty,\]

\[(ii) \quad |\hat{\nu}_i(n) - \nu_i(n)|/n \text{ converges to 0 with probability one.}\]

**Proof:** Observe that \(m_n = m_n\) on \(A_n\). Hence (i) follows from the theorem. The other result is trivial from the structure of \(A_n\). \(\square\)

**Corollary 2.2** Let \(\hat{\mu}_j = (S_{\hat{\nu}_{j+1}} - S_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j), 0 \leq j \leq m_n\), where \(\hat{\nu}_0 = 0\) and \(\hat{\nu}_{m_n+1} = n\). Then, under the conditions of Theorem 2.1, \(\hat{\mu}_j\) converges to \(\mu_j\) in probability.

**Proof:** Observe that

\[S_{\hat{\nu}_{j+1}} - S_{\hat{\nu}_j} = (V_{\hat{\nu}_{j+1}} - V_{\hat{\nu}_j}) + (M_{\hat{\nu}_{j+1}} - M_{\hat{\nu}_j}),\]

where \(M_k\)'s are the partial sums of the means and \(V_k\)'s are the partial sums of the errors. By Lemma 2.1 and the definition of the set \(A_n\), \((V_{\hat{\nu}_{j+1}} - V_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j)\) converges to zero in probability. By the definition of the set \(A_n\), \((M_{\hat{\nu}_{j+1}} - M_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j)\) converges to \(\mu_j\) in probability. \(\square\)

So far we discussed the binary segmentation procedure for the known variance case. This can be generalized to the unknown variance case easily. We use the likelihood ratio statistic for testing the equality of the means of \(Y_1, \ldots, Y_i\) and \(Y_{i+1}, \ldots, Y_d\). This is the two sample \(t\)-statistic given by

\[\tilde{Z}_i = (\bar{Y}_i - \tilde{Y}_i')/\hat{\sigma}_i \sqrt{1/i + 1/(d - i)},\]
where \( \bar{Y}_i = \sum_{j=1}^{i} Y_j / i \) and \( \bar{Y}'_i = \sum_{j=i+1}^{d} Y_j / (d - i) \) are the means of the two samples and \( \hat{\sigma}^2 = \left[ \sum_{j=1}^{i} (Y_j - \bar{Y}_i)^2 + \sum_{j=i+1}^{d} (Y_j - \bar{Y}'_i)^2 \right] / (d - 2) \) is the two sample estimate of the variance. Now we proceed as in the known variance case.

Since the estimate of the variance can be inflated by a large amount by the differences in the mean, it is more difficult to detect the change-points in the unknown variance case. The analysis necessary for the consistency result is also more complex because of the non central \( t \) statistic that is involved when the segment has a change-point. We can overcome some of the difficulties by using an estimate of the variance which converges to the true value at a reasonable rate, for the whole process of segmentation. One such estimate is the average of the square of the consecutive differences \( Y_j - Y_{j-1} \). The problem with this is that it is not independent of the averages \( \bar{Y}_i \) and \( \bar{Y}'_i \) and we cannot use the \( t \) distribution any more. Another way is to use a part of the sample which doesn't have a change-point to estimate the variance and the rest of the sample to detect the change-points and estimate their locations. The consistency result, similar to Theorem 2.1, can be proved for both these approaches by a similar argument, after conditioning on the variance estimate.

### 2.3 A Counter Example

In the previous section we proved the consistency of the binary segmentation procedure under the condition that the change-points are separated by at least \( n^{1-\beta} \) observations, for some \( \beta < 1/8 \). Now we will show by an example that a necessary condition for consistency is that consecutive change-points are separated by at least \( \sqrt{n} \) observations. The following example is one for which the binary segmentation procedure fails asymptotically. We will define the normal random variable \( X_i^{(n)} \) with
mean \( \mu_i^{(n)} \) and unit variance as follows. Let \( k_n = \lfloor n/2 \rfloor, l_n = \lfloor \sqrt{n} \rfloor \) and let

\[
\mu_i^{(n)} = \begin{cases} 
0 & \text{if } 0 < i \leq k_n - l_n \\
1 & \text{if } k_n - l_n < i \leq k_n \\
-1 & \text{if } k_n < i \leq k_n + l_n \\
0 & \text{if } k_n + l_n < i \leq n
\end{cases}
\]

Let \( \hat{m}_n \) be the number of change-points detected in the \( n^{th} \) row by the binary segmentation procedure. We will show that \( P(\hat{m}_n = 0) \) converges to 1 as \( n \) goes to infinity.

Let \( Y_1, \ldots, Y_d \) be independent normal random variables with unit variance. Let \( P_0 \) denote the probability measure under which \( Y_i \)'s have zero mean and let \( P_1 \) denote the probability measure under which \( Y_i \) has mean \( \mu_i^{(d)} \). Let \( Z_i \) be as defined by (2.1). We know that the boundary \( b_d \) should satisfy the condition

\[
P_0(Z_i \leq b_d, 1 \leq i \leq d - 1) \to 1 \text{ as } d \to \infty. \quad (2.28)
\]

Clearly we need \( b_d \to \infty \) as \( d \to \infty \) for (2.28) to hold.

We want to show that

\[
P_1(Z_i \leq b_d, 1 \leq i \leq d - 1) \to 1 \text{ as } d \to \infty. \quad (2.29)
\]

Let \( \lambda_i \) be the mean of \( Z_i \) under \( P_1 \). Then observe that \( \lambda_i \) is non-zero only for \( i \) between \( k_d - l_d \) and \( k_d + l_d \). Also the maximum (\( \eta_d \)) is attained at \( k_d \) by Lemma 2.2 and is \( l_d/[k_d(1 - k_d/d)]^{1/2} \), which is bounded by 3. Since the covariance structure is same under both \( P_0 \) and \( P_1 \), we get that

\[
P_1(Z_i \leq b_d, 1 \leq i \leq d - 1) = P_0(Z_i \leq b_d - \lambda_i, 1 \leq i \leq d - 1). \quad (2.30)
\]

We will show (2.29) by showing that the difference between the probabilities in (2.28) and (2.29) goes to 0. Let \( E_d \) be the difference. Then by (2.30)

\[
E_d = P_0(Z_i \leq b_d, 1 \leq i \leq d - 1) - P_0(Z_i \leq b_d - \lambda_i, 1 \leq i \leq d - 1)
\]
\[
= P_0 \left( \bigcap_{1 \leq i \leq d-1} \{ Z_i \leq b_d \} \cap \{ Z_i > b_d - \lambda_i, \text{ for some } k_d - l_d \leq i \leq k_d + l_d \} \right) \\
\leq P_0 \left\{ \max_{k_d - l_d \leq t \leq k_d + l_d} Z_t > b_d - \eta_d \right\}
\]

Observe that \( \{ Z_i; 1 \leq i \leq d - 1 \} \) is the standardized discrete Brownian Bridge process under \( P_0 \). Let \( W_0 \) be the Brownian Bridge process, \( t_{0d} = (k_d - l_d)/d \) and \( t_{1d} = (k_d + l_d)/d \). Then for large \( d \)

\[
P_0 \left\{ \max_{k_d - l_d \leq t \leq k_d + l_d} Z_t > b_d - \eta_d \right\} \leq P \left\{ \max_{t_{0d} \leq t \leq t_{1d}} \frac{W_0(t)}{\sqrt{t(1-t)}} > b_d - \eta_d \right\} \\
\leq P \left\{ \max_{0.4 \leq t \leq 0.6} \frac{W_0(t)}{\sqrt{t(1-t)}} > b_d - \eta_d \right\}.
\]

The first inequality follows from the fact we are taking maximum over a larger set of random variables. Since both \( t_{0d} \) and \( t_{1d} \) converge to 0.5, the second inequality is true for large \( d \). Since \( \eta_d \) is bounded, \( b_d - \eta_d \) converges to \( \infty \). Finally since \( \max_{0.4 \leq t \leq 0.6} W_0(t)/\sqrt{t(1-t)} \) is finite with probability one, \( P(\bar{m}_n = 0) \) converges to 1, i.e. the binary segmentation procedure fails to detect any change-point.

**Remark 2.3** Observe that in our example we had \( \eta_d \) to be bounded. Instead, if we let \( \eta_d \) go to infinity at a rate lesser than \( b_d \), we will have an example where the minimal distance between change-points increases at a larger rate than \( \sqrt{n} \). Since \( b_d \) increase at a rate at least as large as \( \sqrt{\log \log d} \) we should have minimal distance between change-points to be at least \( \sqrt{n \log \log n} \).
Chapter 3

Schwarz Criterion

In the previous chapter we discussed the binary segmentation procedure. Now we will discuss the other procedure that belongs to the class of methods based on the 'entire' sample. Yao (1988) introduced a procedure using the Schwarz criterion to estimate the number and the locations of changes in normal mean. We will now describe the procedure.

3.1 Description of the Procedure

The idea of the procedure is as follows. For each $m$, we compute the minimal residual sum of squares under the assumption that the sample has $r$ change-points. We then pick the $r$ that minimizes Schwarz criterion, which is a function of the minimal sum of squares and a cost per change-point. More precisely, let $X_1, \ldots, X_n$ be the data and let $S_k = X_1 + \cdots + X_k$. Let $0 = i_0 < i_1 < \cdots < i_r < i_{r+1} = n$ be a partition, i.e., the $r$ hypothesized change-points. This partitions the data into $r + 1$ groups, $\{X_{i_{j+1}}, \ldots, X_{i_{j+1}}\}$, $0 \leq j \leq r$. Let $SS(r, n)$ denote the minimal residual sum of squares, that is the minimum of the residual sum of squares over all such partitions.
Then the Schwarz criterion is defined as

\[ SC(r, n) = (n/2) \log(SS(r, n)/n) + \alpha_n r \log n, \]

where \( \{\alpha_n\} \) is a sequence of constants denoting the cost per change-point. We will estimate the number of change-points \( (m_n) \) by the \( r \) that minimizes the Schwarz criterion. The partition that gives us the minimal residual sum of squares gives us the estimates of the locations of the change-points, that is, \( \hat{v}_j = i_j \). Unfortunately the procedure requires an enormous amount of numerical calculation unless \( r \) is quite small (say \( r \leq 3 \)). In his paper, Yao sets the cost \( \alpha_n \) to be 1. We will use a different sequence which will be derived as a part of our consistency theorem.

**Remark 3.1** Observe that for \( r = 1 \), the minimal residual sum of squares is attained at the same \( i \) where the likelihood ratio statistic is maximized. So both this procedure and the binary segmentation procedure give the same estimate of the location of the change-point.

There are at least two ways of estimating the number of change-points. One way is to compute the Schwarz criterion for all \( r \) up to \( R_n \), where \( R_n \) is an upper bound for the number of change-points, which in our case will be about \( n/\log n \), and take the minimum. The other way is to compute \( SC \) sequentially over \( r \) and stop soon after it starts increasing. We will take \( r \) as the first time \( SC(r, n) \) is smaller than \( SC(r + 1, n) \), \( SC(r + 2, n) \) and \( SC(r + 3, n) \). The sequential method is preferable because the amount of computation increases at an incredible rate with \( r \).

Observe that \( SC(r, n) - SC(r + 1, n) \) is invariant under scale transformation of the data (because both \( SS(r, n) \) and \( SS(r + 1, n) \) are multiplied by the same constant). So the estimates of the number of change-points and their locations do not require the knowledge of the variance \( \tau^2 \). So the procedure works both in the known and the unknown variance case. In the next section we will derive a consistency theorem for this procedure.


3.2 Consistency Result

Yao proves the consistency of his procedure when there are a fixed number of change-points, say \( m \). He needs that an upper bound for \( m \) is known and the locations of the change-points satisfy the condition \( \nu_j^{(n)} / n \) converges to \( p_j \) where \( 0 < p_1 < \ldots < p_m < 1 \). We will now derive a consistency result under weaker conditions. From the observation in the previous section, we can assume without loss of generality that the variance is known and is 1.

First we will prove the following preliminary results. Let \( U_1, U_2, \ldots \) be independent standard normal random variables and let \( V_k = U_1 + \cdots + U_n \) be their partial sums.

**Lemma 3.1** Fix \( \eta > 0 \) and let

\[
B_n = \left\{ \max_{1 \leq i < j \leq n} \frac{(V_j - V_i)^2}{j - i} \leq 2(1 + \eta) \log n \right\}. \tag{3.1}
\]

Then \( P(B_n^c) \to 0 \) as \( n \to \infty \).

**Proof:** For \( 1 \leq i \leq n \) let

\[
B_{i,n} = \left\{ \max_{1 \leq j \leq n-i+1} \frac{V_j}{\sqrt{j}} > \sqrt{2(1 + \eta) \log n} \right\}.
\]

Then observe that

\[
P(B_n^c) \leq 2 \sum_{i=1}^{n} P(B_{i,n}) \leq 2nP(B_{1,n}). \tag{3.2}
\]

We will now derive an upper bound for \( P(B_{1,n}) \).

Let \( \{W_t, 0 \leq t < \infty\} \) be the standard Brownian motion and let \( \tau \) be the stopping time defined by

\[
\tau = \inf \left\{ t : t \geq 1 \text{ and } W_t \geq b\sqrt{t} \right\}.
\]

Then we have

\[
P \left( \max_{1 \leq i \leq n} \frac{V_i}{\sqrt{i}} \geq b \right) \leq P \left( \max_{1 \leq i \leq n} \frac{W_t}{\sqrt{t}} \geq b \right) = P(\tau \leq n).
\]
We will find an upper bound for the probability on the right by approximating the square root boundary by piecewise linear boundaries. Let $1 = t_0 < t_1 < \ldots < t_m = n$ be a partition of $[1, n]$ and let $t_{-1} = 1$. For $0 \leq j \leq m$, let $\tau_j$ be the stopping time given by

$$
\tau_j = \inf \{ t : W_t \geq \alpha_j + \beta_j t \}
$$

where $\alpha_j = \frac{b \sqrt{t_{j-1} - t_j}}{\sqrt{t_{j-1} + \sqrt{t_j}}}$ and $\beta_j = \frac{b}{\sqrt{t_{j-1} + \sqrt{t_j}}}$. Observe that for all $1 \leq j \leq m$

$$
\{t_{j-1} \leq \tau \leq t_j\} \subset \{t_{j-1} \leq \tau_j \leq t_j\} \cup \{\tau_0 \leq 1\}.
$$

So we get that

$$
\{\tau \leq n\} \subset \{\tau_0 \leq 1\} \cup \bigcup_{1 \leq j \leq m} \{t_{j-1} < \tau_j \leq t_j\}.
$$

Let $T$ be the stopping time given by $T = \inf \{ t : W_t > \alpha + \beta t \}$. Then the density of $T$ is given by $\frac{\beta}{t^{1/2}} \phi(\beta t^{-1/2} + \alpha t^{1/2})$ (e.g. Siegmund (1985)). From this we get that

$$
P(\tau_0 \leq 1) = 1 - \Phi(b) + e^{-bt^2/2}/2
$$

and for $1 \leq j \leq m$

$$
P(t_{j-1} \leq \tau_j \leq t_j) = \int_{t_{j-1}}^{t_j} \frac{b \sqrt{t_{j-1} - t_j}}{(\sqrt{t_{j-1} + \sqrt{t_j}})^{3/2}} \phi\left(\frac{b \sqrt{t_{j-1} - t_j}}{(\sqrt{t_{j-1} + \sqrt{t_j}})^{1/2}} + \frac{bt^{1/2}}{\sqrt{t_{j-1} + \sqrt{t_j}}}\right) dt.
$$

Summing these over $j$ and taking limit as partition size converges to zero we get that

$$
P(\tau \leq n) \leq P(\tau_0 \leq 1) + \sum_{1 \leq j \leq m} P(t_{j-1} \leq \tau_j \leq t_j)

\rightarrow 1 - \Phi(b) + e^{-bt^2/2}/2 + \int_1^n \frac{b}{2t} \phi(b) dt

\leq 1 - \Phi(b) + b\phi(b)(\log n + 1).

(3.3)

Now by substituting the value $\sqrt{2(1 + \eta)\log n}$ for $b$ in the bound given by (3.3) and using it in equation (3.2) we can see that $P(E_n^c) \rightarrow 0$ as $n \rightarrow \infty$. \(\square\)

Lemma 3.1 is the same as Lemma 1 in Yao (1988) but the proof given here is different from the one in Yao. This proof uses the boundary crossing argument for standard Brownian motion (e.g. Ito and McKeon (1965 p. 34)).
Lemma 3.2 On $B_n$ $\max_{1 \leq i < j \leq n} \frac{(k-j)(j-i)}{(k-i)} \left( \frac{V_k - V_j}{k-j} - \frac{V_j - V_i}{j-i} \right)^2 \leq 4(1 + \eta) \log n$.

Proof: Use the definition of $B_n$ as follows

$$\frac{(k-j)(j-i)}{(k-i)} \left( \frac{V_k - V_j}{k-j} - \frac{V_j - V_i}{j-i} \right)^2 \leq \frac{(k-j)(j-i)}{(k-i)} \left( \frac{\sqrt{2(1 + \eta) \log n}}{\sqrt{k-j}} + \frac{\sqrt{2(1 + \eta) \log n}}{\sqrt{j-i}} \right)^2 \leq 4(1 + \eta) \log n.$$ 

We will now prove the following consistency theorem. We need the following definitions. Let $\delta_j = \theta_j - \theta_{j-1}$, $1 \leq j \leq m_n$. Let $\pi = (i_1, \ldots, i_r)$ be a partition of size $r$. Let $\gamma_j = \#\{l : \nu_j < i_l \leq \nu_{j+1}\}, j = 0, \ldots, m_n$.

Theorem 3.1 Let $\{X_i^{(n)}; 1 \leq i \leq n \leq \infty\}$ be a Gaussian process defined by (1.2). Assume that the following conditions hold:

(i) $\inf_j |\nu_{j+1}^{(n)} - \nu_j^{(n)}| \geq c_n \log n$ for some $c_n \to \infty$ as $n \to \infty$

(ii) $\inf_j |\theta_{j+1} - \theta_j| \geq \delta$ for some $\delta > 0$ independent of $n$

(iii) $\sup_j |\theta_i| \leq B$ for some $B$ independent of $n$.

Let the Schwarz criterion be defined using an appropriately chosen cost function $\{c_n\}$. The conditions on the cost function is derived in the proof. Let $(i_1, \ldots, i_r)$ be a partition that minimizes the Schwarz criterion. Let $c'_n$ be such that $c'_n \leq c_n/2$, $c'_n$ goes to infinity and $c'_n/c_n$ converges to zero. Let

$$A_n = \{r = m, |\nu_j - i_j| \leq c'_n \log n, 1 \leq j \leq m_n\} \quad (3.4)$$

Then $P(A_n)$ converges to 1 as $n$ goes to $\infty$.

Proof: We will prove that $B_n$ is contained in $A_n$ and so by Lemma 3.1 the theorem follows. Let $U_i = X_i - \mu_i$, $1 \leq i \leq n$. From now we will restrict ourselves to the $U$'s contained in $B_n$. and we will no longer use the superscript $(n)$ unless it is necessary.
CHAPTER 3. SCHWARZ CRITERION

From the definition of Schwarz Criterion we get the following two equations

\[ SC(r, n) - SC(r + 3, n) = \frac{n}{2} \log \left( 1 + \frac{SS(r, n) - SS(r + 3, n)}{SS(r + 3, n)} \right) - 3\alpha_n \log n \]  
(3.5)

\[ SC(r - 1, n) - SC(r, n) = \frac{n}{2} \log \left( 1 + \frac{SS(r - 1, n) - SS(r, n)}{SS(r, n)} \right) - \alpha_n \log n \]  
(3.6)

By Lemma 3.6, for \( r < m_n \), \( SS(r, n) - SS(r + 3) \geq \kappa c_n \log n \) and \( SS(r + 3, n) \leq n(2 + B^2) \) by Lemmas 3.3 and 3.4. Using these in (3.5), for large \( n \), we get

\[ SC(r, n) - SC(r + 3, n) \geq \frac{n}{2} \log \left( 1 + \frac{\kappa c_n \log n}{n(2 + B^2)} \right) - 3\alpha_n \log n \]  
(3.7)

Similarly for \( r > m_n \), since \( SS(r - 1, n) - SS(r, n) \leq 120(1 + \eta) \log n \) by Lemma 3.7 and \( SS(r, n) \geq n(1 - \epsilon) \) by Lemma 3.4, for large \( n \), we get from equation (3.6)

\[ SC(r - 1, n) - SC(r, n) \leq \frac{n}{2} \log \left( 1 + \frac{120(1 + \eta) \log n}{n(1 - \epsilon)} \right) - \alpha_n \log n \]  
(3.8)

In order to make the right hand side of (3.7) larger than zero and the right hand side of (3.8) smaller than zero we should take \( \alpha_n \) such that for large \( n \) it is larger than \( 60(1 + \eta)/(1 - \epsilon) \) but should grow at a smaller rate than \( c_n \). Since we don’t know \( c_n \) we can set \( \alpha_n = 65 \) by choosing \( \eta \) and \( \epsilon \) suitably. Thus we get that the Schwarz criterion is minimized for \( r = m_n \) over a set with probability converging to 1.

Now we have to prove that \(|\nu_j - i_j| \leq c'_n \log n\) for all \( j \). We will prove this by contradiction. Suppose there exists a \( j \) such that it is not true. Then, by (3.10) of Lemma 3.5, we get that we can reduce the sum of squares by at least \( \delta^2 c'_n \log n/8 \)

This says that \( SS(m_n, n) - SS(m_n + 3, n) \) goes to infinity at a rate larger than \( c'_n \log n \). But, by Lemma 3.7, \( SS(m_n, n) - SS(m_n + 3, n) \) has a rate bounded by \( 360(1 + \eta) \log n \). Since the two statements are contradicting one another we get the theorem. Q.E.D.

**Lemma 3.3** Under the assumptions of the theorem \( P\{SS(0, n) \leq n(B^2 + 2)\} \) converges to 1 as \( n \) goes to infinity.
CHAPTER 3. SCHWARZ CRITERION

Proof: Since $SS(0, n)$ is the total sum of squares it follows a noncentral chi-squared distribution with $n - 1$ degrees of freedom and noncentrality parameter $\lambda$ given by $\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2$ where $\bar{\mu}$ is the average of the $\mu_i$'s. By assumption (iii) of the theorem $\lambda$ is bounded by $nB^2$. Observe that $SS(0, n)$ can be written as the sum of independent random variables $U$ and $V$ where $U$ has a central chi-squared with $n - 2$ degrees of freedom and $V$ is the square of a $N(\sqrt{\lambda}, 1)$ random variable. So

$$P\{SS(0, n) \leq n(B^2 + 2)\} \geq P\{U \leq 3n/2\}P\{V \leq n(B^2 + 1/2)\}. \quad (3.9)$$

Since the right hand side of (3.5) converges to 1, the result follows. \qed

Lemma 3.4 Let $p_n = n/c_n \log n$ and let $m_n \leq r \leq p_n$. Then, as $n$ goes to infinity, for all $\epsilon$ positive, $P\{SS(r, n)/n \in (1 - \epsilon, 1 + \epsilon)\}$ converges to 1.

Proof: Since $SS(r, n)$ is monotonically decreasing in $r$ it is enough to get an upper bound for $SS(m_n, n)$ and a lower bound for $SS(p_n, n)$. For the upper bound observe that $SS(m_n, n)$ is less than the residual sum of squares we get if we partition the data set at the change-points. This residual sum of squares, which we shall denote by $W$, follows a chi-squared distribution with $n - m_n - 1$ degrees of freedom. Thus by the law of large numbers we get the upper bound.

As for the lower bound, let $(i_1, \ldots, i_{p_n})$ be a partition that corresponds to the minimal sum of squares $SS(p_n, n)$. If we add the change-points to the partition we reduce the sum of squares. Call the new sum of squares $W'$. Observe that we can get $W'$ from $W$ by removing sum of squares within each group of data with constant mean. By Lemma 3.2 each such number is bounded by $4(1 + \eta)\log n$. By the law of large numbers $W$ is bounded below by $n(1 - \epsilon/2)$ with probability converging to 1. Since $p_n \log n$ converges to zero we get the result. \qed

Lemma 3.5 Let $n_1$ and $n_2$ be such that, $\nu_{j-1} < n_1 < \nu_j < n_2 < \nu_{j+1}$ and $|\nu_j - n_i| \geq
\[32(1 + \eta) \log n/\delta_j^2 \quad (i = 1, 2). \text{ Then} \]
\[
\frac{(n_2 - \nu_j)(\nu_j - n_1)}{(n_2 - n_1)} \left( \frac{S_{n_2} - S_{\nu_j}}{n_2 - \nu_j} - \frac{S_{\nu_j} - S_{n_1}}{\nu_j - n_1} \right)^2 \geq 4(1 + \eta) \log n.
\]

**Proof:** Let \( k_i = |\nu_j - n_i|, i = 1, 2. \) Without loss of generality assume that \( \mu_{j+1} = \delta_j \) and \( \mu_j = 0 \) and that \( \delta_j > 0. \) Then
\[
\frac{(n_2 - \nu_j)(\nu_j - n_1)}{(n_2 - n_1)} \left( \frac{S_{n_2} - S_{\nu_j}}{n_2 - \nu_j} - \frac{S_{\nu_j} - S_{n_1}}{\nu_j - n_1} \right)^2 \\
\geq \frac{k_1k_2}{k_1 + k_2} \left( \delta_j - \sqrt{2(1 + \eta) \log n/k_1} - \sqrt{2(1 + \eta) \log n/k_2} \right)^2 \\
\geq \frac{k_1k_2}{k_1 + k_2} (\delta_j/2)^2 \\
\geq \delta_j^2 (k_1 \land k_2)/8 \quad (3.10)
\]

The first inequality is from Lemma 3.1 and the second from the conditions on \( k_1 \) and \( k_2. \) The third inequality is simple algebra. Since \( k_1 \land k_2 \) is bounded below by \( 32(1 + \eta) \log n/\delta_j^2 \) we get the lemma. \( \square \)

**Lemma 3.6** For \( r < m_n, SS(r, n) - SS(r + 3, n) \geq \kappa c_n \log n, \) for large \( n, \) where \( \kappa \) is a constant independent of \( n. \)

**Proof:** Let \((i_1, \ldots, i_r)\) correspond to \( SS(r, n). \) Since \( r \) is less than \( m_n, \) by condition \((i)\) of the theorem, there exists \( j \) such that \( i_l \notin (\nu_j - c_n \log n/2, \nu_j + c_n \log n/2) \) for all \( l. \) Let \( j_0 \) be one such \( j. \) Let \( n_1 = \nu_{j_0} - c_n \log n/2, \) \( n_2 = \nu_{j_0} \) and \( n_3 = \nu_{j_0} + c_n \log n/2. \) The new partition we get by adding \( n_1, n_2 \) and \( n_3 \) to the \( i \)'s decreases the residual sum of squares by at least \([n_3(n_2 - n_1)]((S_{n_3} - S_{n_2})/(n_3 - n_2) - (S_{n_2} - S_{n_1})/(n_2 - n_1))^2. \) For large \( n,
\[
\frac{(n_3 - n_2)(n_2 - n_1)}{(n_3 - n_1)} \left( \frac{S_{n_3} - S_{n_2}}{n_3 - n_2} - \frac{S_{n_2} - S_{n_1}}{n_2 - n_1} \right)^2 \geq \delta_j^2 c_n \log n/16.
\]

The inequality follows form (3.10) in Lemma 3.5 by setting \( j = j_0 \) and and using \( \delta_j^2 \geq \delta_j^2 \) for all \( j. \) Since \( SS(r + 3, n) \) is smaller than the residual sum of squares of this partition of size \( r + 3, \) we get the lemma. \( \square \)
Lemma 3.7 For \( r > m_n \), \( SS(r-1,n) - SS(r,n) \leq 120(1+\eta) \log n \) for large \( n \).

**Proof:** Let \( \pi = (i_1, \ldots, i_r) \) be a partition of size \( r \) with minimal residual sum of squares. By Lemma 3.2, if \( \gamma_0 \geq 2 \) or \( \gamma_{m_n} \geq 2 \) or \( \gamma_j \geq 3 \) for \( 0 < j < m_n \) the result follows. So we can assume without loss of generality that \( \gamma_0 \) and \( \gamma_{m_n} \) are at most 1 and the other \( \gamma_j \)'s are less than 2. Since \( \sum_{j=0}^{m_n} \gamma_j = r \) is at least \( m_n + 1 \), \( \gamma_j \)'s satisfy one of the following conditions. (1) \( \gamma_j \geq 1 \) for all \( j \), (2) there exists a \( k \) such that \( \gamma_k = 2 \) and \( \gamma_j = 1 \) either for all \( j < k \) or for all \( j > k \) and (3) there exist \( k \) and \( l \) with \( k < l \) such that \( \gamma_k = \gamma_l = 2 \) and \( \gamma_j = 1 \) for all \( k < j < l \). The result, for all three cases, follows from the next lemma. \( \square \)

Lemma 3.8 Let \( Y_1, \ldots, Y_d \) (\( d \leq n \)) be the data and let \( 0 = \zeta_0 < \zeta_1 < \ldots < \zeta_h < \zeta_{h+1} = d \) be the change points. Let \( \zeta_i - \zeta_{i-1} \geq c_n \log n, 2 \leq i \leq h \). Let \( U_i = Y_i - E(Y_i) \) and \( V_i = U_1 + \cdots + U_i \). For all \( 0 \leq i < j < k \leq d \), let

\[
\frac{(k-j)(j-i)}{(k-i)} \left[ \frac{(V_k - V_j)}{(k-j)} - \frac{(V_j - V_i)}{(j-i)} \right]^2 \leq 4(1+\eta) \log n.
\]

Let \( 0 < i_1 < \zeta_1 < i_2 < \zeta_2 < \ldots < i_h < \zeta_h < i_{h+1} < d \) be a partition of size \( h + 1 \). Then there exists a partition \( (j_1, \ldots, j_h) \) such that the difference in the residual sum of squares of the two partitions is at most \( 120(1+\eta) \log n \).

**Proof:** Let \( \omega_j \) indicate the change in mean at \( \zeta_j \). Let \( l_1 \) be the first \( l \) such that \( i_{l+1} - \zeta_l \leq 32(1+\eta) \log n/\omega_l^2 \). Set \( l = h + 1 \) if no such \( l \) exists. Let \( j_0 \) be the largest \( l \) less than \( l_1 \) that satisfies \( \zeta_l - i_l \leq 32(1+\eta) \log n/\omega_l^2 \). Let \( l_0 = 0 \) if no such \( l \) exists. By Lemma 3.5, by adding \( \zeta_i, l_0 < l < l_1 \), to the partition we decrease the residual sum of squares by at least \( 4(1+\eta) \log n \), for each point added. Now if we remove \( i_i, l_0 + 1 < l < l_1 \), from the new partition we increase the sum of squares by at most \( 4(1+\eta) \log n \), for each point deleted. Finally, by Lemma 3.9, removing \( i_{l_0+1} \) and \( i_i \) from the partition increases the sum of squares by at most \( 60(1+\eta) \log n \), for each
point deleted. Thus we have a new partition of size $h$ such that the sum of squares of the new one differs from the old partition by at most $120(1 + \eta) \log n$. □

**Lemma 3.9** Let $Y_1, \ldots, Y_d$ ($d \leq n$) be the data and let $0 = \zeta_0 < \zeta_1 < \zeta_2 < \zeta_3 = d$ be the change points. Let $\omega_1$, 0 and $\omega_2$ be the means of the $Y$'s in the range $\{1, \ldots, \zeta_1\}$, $\{\zeta_1 + 1, \ldots, \zeta_2\}$ and $\{\zeta_2 + 1, \ldots, \zeta_3\}$ respectively. Let $\zeta_2 - \zeta_1 \geq c_n \log n$, $\zeta_1 \leq 32(1 + \eta) \log n/\omega_2^2$ and $d - \zeta_2 \leq 32(1 + \eta) \log n/\omega_2^2$. Let $U_i = Y_i - E(Y_i)$ and $V_i = U_1 + \cdots + U_i$. For all $0 \leq i < j < k \leq d$, let

\[
\frac{(k - j)(j - i)}{(k - i)} \left( \frac{(V_k - V_j)}{(k - j)} - \frac{(V_j - V_i)}{(j - i)} \right)^2 \leq 4(1 + \eta) \log n
\]

Let $\zeta_1 < l < \zeta_2$. Then the sum of squares between the two groups is at most $60(1 + \eta) \log n$.

**Proof:** The sum of squares between the two groups can be written as $(a + b)^2$, where $a^2 = [l(d - l)/d][E(V_i)/l - (E(V_d) - E(V_i))/(d - l)]^2$ and $b^2 = [l(d - l)/d][V_i/l - (V_d - V_i)/(d - l)]^2$. By the conditions on $\zeta_1$ and $\zeta_2$, $a^2$ is bounded by $32(1 + \eta) \log n$ and by the hypothesis in the lemma, $b^2$ is bounded by $4(1 + \eta) \log n$. So $(a + b)^2$ is bounded by $60(1 + \eta) \log n$. □

**Remark 3.2** We set the cost per change-point $\alpha_n$ to be $60 \log n$. By Lemma 3.1 we can see that if the cost were less than $\log n$ we will detect change-points when there aren't any. So it is necessary to have a cost of at least $\log n$. Yao proved that, for his problem, it is sufficient. It remains to be seen if it is sufficient for our problem too.

We can derive the following corollaries from the theorem.

**Corollary 3.1** Under the conditions of Theorem 2.1 the following are true.

(i) $P(\hat{n}_m = m_n) \rightarrow 1$ as $n \rightarrow \infty$,

(ii) $|\hat{\nu}_i^{(n)} - \nu_i^{(n)}|/n$ converges to 0 with probability one.
Chapter 3. Schwarz Criterion

Proof: The results follow from the definition of $A_n$ and the theorem.

Corollary 3.2 Let $\hat{\mu}_j = (S_{\hat{\nu}_{j+1}} - S_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j)$, $0 \leq j \leq \hat{\nu}_n$, where $\hat{\nu}_0 = 0$ and $\hat{\nu}_{\hat{\nu}_n+1} = n$. Then, under the conditions of Theorem 2.1, $\hat{\mu}_j$ converges to $\mu_j$ in probability.

Proof: Observe that

$$S_{\hat{\nu}_{j+1}} - S_{\hat{\nu}_j} = (V_{\hat{\nu}_{j+1}} - V_{\hat{\nu}_j}) + (M_{\hat{\nu}_{j+1}} - M_{\hat{\nu}_j}),$$

where $M_k$'s are the partial sums of the means and $V_k$'s are the partial sums of the errors. By Lemma 3.1 and the definition of the set $A_n$, $(V_{\hat{\nu}_{j+1}} - V_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j)$ converges to zero in probability. By the definition of the set $A_n$, $(M_{\hat{\nu}_{j+1}} - M_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j)$ converges to $\mu_j$ in probability.
Chapter 4

Pseudo-Sequential Procedures

We saw that the estimates from the methods in the previous chapter depended on the entire sample. We also saw in the counter example how the effects of successive changes cancelled one another and prevented us from detecting the changes. We will now discuss the class of pseudo-sequential procedures which are sequential (online) schemes that have been adapted for the fixed sample case. The pseudo-sequential methods are motivated by the idea that the data between the \( j - 1 \)st and \( j + 1 \)st change-points should be the most important factors in the detection and estimation of the \( j \)th change-point. We will be concerned with the pseudo-sequential methods in this chapter. We will now describe the general structure of the pseudo-sequential procedures.

4.1 Description of the Procedure

The basic idea of this class of methods is as follows. We use a sequential method, say for example the CUSUM test, to test for a change-point in the data. Every time we detect a change-point we start all over again with the rest of the data that has not been used so far. We define the successive pseudo-stopping rules for this scheme as
follows. Let \( N_0^{(n)} = 0 \). Once we have defined \( N_0^{(n)}, \ldots, N_i^{(n)} \) we define \( N_{i+1}^{(n)} \) by

\[
N_{i+1}^{(n)} = \inf \left\{ k : N_i^{(n)} < k \leq n, S \left( X_{N_i^{(n)}+1}^{(n)}, \ldots, X_k^{(n)} \right) \geq b_n \right\}, \quad (4.1)
\]

where \( S \) is a suitably chosen statistic and \( b_n \) a suitably chosen boundary. We then estimate the number and the locations of the change-points as follows:

\[
m_n = \sup \left\{ k : N_k^{(n)} < \infty \right\} \quad \text{and} \quad \hat{\nu}_i^{(n)} = N_i^{(n)}, 1 \leq i \leq m_n. \quad (4.2)
\]

As mentioned in Section 1.3, the estimates from this procedure depend on the order in which the data has been used. One way to overcome this problem would be to run the procedure on the data in both order (i.e. 1 to \( n \) and \( n \) to 1) and take a function of the estimates, say average. From the consistency theorem for the estimates given by equation (4.2) we could see that for most such functions the estimates would still be consistent. But we may encounter other problems, such as, what would be the estimate of the number of change-points when the two estimates of the number of change-points are not the same ?. So we will use the estimates we get by using the procedure on the data in the order 1 to \( n \).

We will now discuss a prototype of the consistency theorems for the pseudo-sequential procedures.

### 4.2 General Consistency Statement

In this section we will describe the general consistency statement for the pseudo-sequential methods and the method of proof. This consistency statement is a prototype of the consistency results we derive in Chapter 3.

**Statement 4.1** Let \( \{ X_i^{(n)} ; 1 \leq i \leq n < \infty \} \) be an array of random variables as described in the general statement of the problem satisfying equation 1.1. Assume that
the following condition on the minimal distance between change-points holds
\[ \inf_i |\nu_{i+1}^{(n)} - \nu_i^{(n)}| \geq 2\beta_n, \text{ for all } n, \quad (4.3) \]
where \( \{\beta_n\} \) is a sequence of constants which depends on \( \{G_j\} \). Let
\[ \mathcal{A}_n = \left\{ \nu_i^{(n)} < N_i^{(n)} < \nu_i^{(n)} + \beta_n, 1 \leq i \leq m_n \& N_{m_n+1}^{(n)} = \infty \right\}. \quad (4.4) \]

Then under suitable conditions on the boundary \( b_n \), the lower bound on the minimal distance between change-points \( \beta_n \) and the class of distribution functions \( \{G_j\} \), which are part of the individual theorems, we can prove that \( \text{Prob}(\mathcal{A}_n) \) converges to 1 as \( n \) goes to infinity.

**Proof:** The logical structure of a proof of a theorem of this type will be described here. We need the following setup. Let \( \{Y_i\} \) be a sequence of independent random variables. Let \( P_{\nu}^{(j)} \) denote the probability measure under which \( Y_1, \ldots, Y_\nu \) are i.i.d. with distribution function \( G_j \) and \( Y_{\nu+1}, \ldots \) are i.i.d. with distribution function \( G_{j+1} \). Let \( N \) denote the stopping time
\[ N = \inf \{k : S(Y_1, \ldots, Y_k) \geq b\}. \quad (4.5) \]

Observe that
\[ \mathcal{A}_n^c \subset B_{0n} \cup \bigcup_{i=1}^{m_n} \{B_{in} \cup C_{in}\} \quad (4.6) \]
where
\[ B_{in} = \left\{ \nu_i^{(n)} < N_i^{(n)} < N_{i+1}^{(n)} \leq \nu_{i+1}^{(n)} \right\}, \quad 0 \leq i \leq m_n, \text{ and } \]
\[ C_{in} = \left\{ \nu_j^{(n)} < N_j^{(n)} < \nu_j^{(n)} + \beta_n, \ 1 \leq j < i \& N_i^{(n)} > \nu_i^{(n)} + \beta_n \right\}, \quad 1 \leq i \leq m_n \quad (4.8) \]

We will now get upper bounds for the probabilities \( P(B_{in}) \) and \( P(C_{in}) \). Observe that
\[ P(B_{in}) = E\left\{ P\left( \nu_i^{(n)} < N_i^{(n)} < N_{i+1}^{(n)} \leq \nu_{i+1}^{(n)} | N_i^{(n)} \right) \right\} \]

= E \left\{ P \left( N_{i+1}^{(n)} - N_i^{(n)} \leq \nu_{i+1}^{(n)} - N_i^{(n)}|N_i^{(n)} \right) \right\} \\
\leq E \{ P_{\infty}^{(i)}(N \leq n) \} \\
= P_{\infty}^{(i)}(N \leq n).

The first, second and the final equality are trivial. The inequality follows from the fact that $N_i^{(n)}$ and $N_{i+1}^{(n)} - N_i^{(n)}$ are independent and $\nu_{i+1}^{(n)} - N_i^{(n)} \leq n$. Similarly

$$P(C_{in}) \leq E \left\{ P \left( N_i^{(n)} > \nu_i^{(n)} + \beta_n|N_i^{(n)} \right) \right\}$$

$$= E \left\{ P \left( N_i^{(n)} - N_{i-1}^{(n)} > \left( \nu_i^{(n)} - N_{i-1}^{(n)} \right) + \beta_n|N_i^{(n)} \right) \right\}$$

$$\leq E \left\{ \max_{\beta_n \leq \nu \leq n-\beta_n} P_{\nu}^{(i-1)}(N > \nu + \beta_n) \right\}$$

$$= \max_{\beta_n \leq \nu \leq n-\beta_n} P_{\nu}^{(i-1)}(N > \nu + \beta_n).$$

As before the first inequality and the equalities are trivial. The second inequality follows from the facts that $N_{i-1}^{(n)}$ and $N_i^{(n)} - N_{i-1}^{(n)}$ are independent and by equations (4.3) and (4.8), $\nu$ satisfies $\beta_n \leq \nu \leq n - \beta_n$, where $\nu = \nu_i^{(n)} - N_{i-1}^{(n)}$.

Let

$$p_{1i}^{(n)} = P_{\infty}^{(i)}(N \leq n) \quad \text{and} \quad p_{2i}^{(n)} = \max_{\beta_n \leq \nu \leq n-\beta_n} P_{\nu}^{(i-1)}(N > \nu + \beta_n). \quad (4.9)$$

If the class of distribution functions $\{G_j\}$ and the choice of $b_n$ and $\beta_n$ are such that $p_1$ and $p_2$ satisfy the condition

$$\sum_{i=0}^{m} p_{1i}^{(n)} + \sum_{i=1}^{m} p_{2i}^{(n)} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty,$$

then $P(A_n^c) \rightarrow 0$ as $n \rightarrow \infty$ and the proof of the theorem follows.

\[\Box\]

**Remark 4.1** The two sequences $p_{1i}^{(n)}$ and $p_{2i}^{(n)}$ give us bounds for the error probabilities of two types. While the former gives bounds for probability of detecting a change when there isn't any, the latter bounds the probability of taking a large number of observations to detect a change once the change has occurred. There is a trade off between the two types of errors i.e. if we try to reduce the error probability of one type we end up increasing the error probability of the other type.
CHAPTER 4. PSEUDO-SEQUENTIAL PROCEDURES

We can prove this consistency result for various pseudo-sequential procedures by applying Statement 4.1 and the method of proof. We can also derive various corollaries, like the consistency of the estimated number and locations of the change-points from this result.

4.3 Special Cases

We will study the problem of multiple changes-points in the mean function of a Gaussian process with constant variance as given by (1.2). We will be dealing with two types of procedures depending on the motivation behind the statistic $S$ used in defining the pseudo-stopping rules in (4.1). If $S$ is motivated by Bayesian arguments (as in the Shiryaev-Roberts scheme) we will call them Bayesian procedures. If it is derived using maximum likelihood considerations (as in CUSUM tests) we will call them likelihood ratio procedures. We will discuss these two in detail in Chapters 5 and 6 respectively. We will derive appropriate consistency results for both types by applying the general consistency statement given in Section 4.2. In Chapter 8 we will also study the case of exponential processes. We will derive the Bayesian procedure and prove a consistency theorem for the procedure in Section 8.1. Numerical results for the two cases are in Chapter 7 and Section 8.2 respectively.

Since we need the following set of definitions and notations throughout the next two chapters, we will describe it here and refer to it whenever it is necessary.

[*] Let $Y_1, Y_2, \ldots$ be independent normal random variables with unit variance. Let $P_\infty$ indicate the probability measure under which all the $Y_i$'s have zero mean and $P_\nu^{(u)}$ the measure under which $Y_1, \ldots, Y_\nu$ have mean zero and $Y_{\nu+1}, Y_{\nu+2}, \ldots$ have mean $\mu$. Let $\tilde{S}_k = Y_1 + \cdots + Y_k$. Let $N$ be the stopping time defined by

$$N = \inf\{k : S(Y_1, \ldots, Y_k) > b\},$$

where the statistic $S$ and the boundary $b$ are suitably defined. Let $B_i, 0 \leq i \leq m_n,$
and $C_i, 1 \leq i \leq m_n$, be as defined in (4.7) and (4.8) respectively. From the consistency statement and the method of proof in Section 4.2, we have to find good upper bounds for $P_\infty(N \leq n)$ and
\[
\max_{\beta_n \leq \nu \leq n-\beta_n} P^{(\nu)}(N > \nu + \beta_n)
\]
in order to prove the consistency theorem.
Chapter 5

Bayesian Method

In this chapter we will discuss the Bayesian method. These are pseudo-sequential procedures where the test criteria are derived through Bayesian arguments. They are derived from the methodology developed by Pollak and Siegmund (1991) to detect a change in the mean of normal random variables when the initial mean is unknown. The variance of the Gaussian process defined in (1.2) can either be known or unknown. We will discuss the known variance case in detail in Sections 5.1 and 5.2. Finally in Section 5.3 we will derive the test statistic for the unknown variance case and talk about the difficulties involved in analyzing them.

5.1 Derivation of the Statistic

Let \( X_1, X_2, \ldots \) be independent normal random variables with unit variance and let \( S_k = X_1 + \cdots + X_k \). Motivated by Bayesian arguments of Shiryaev (1963) in the case the initial value of the mean is known and by considerations of invariance to handle the case when the initial mean value is unknown, Pollak and Siegmund studied the
statistic
\[ R_k(\delta) = \sum_{i=0}^{k-1} \exp \left[ \delta(iS_k/k - S_i) - \delta^2 i(1 - i/k)/2 \right], \] (5.1)

where \( \delta > 0 \) is a parameter chosen by the experimenter, to detect a change in the mean. We now derive give an alternative derivation of the statistic (5.1) by using the the likelihood ratio based on the invariant functions of the observations and a uniform prior for the location of the change-points.

Since the problem of detecting a change-point is location invariant, we would like to derive a location invariant test. The location invariant transformation of the data is
\[ Y_1 = X_1, \ Y_j = X_j - X_1, \ 2 \leq j \leq n. \]

We want to test the hypothesis of no change against the alternative that there is a change in the mean at \( k \) and the difference in the means is \( \theta \). We will use the likelihood ratio statistic based on the location invariant transformation of the data \( \{Y_j, 2 \leq j \leq n\} \). The inverse transformation is given by
\[ X_1 = Y_1, \ X_j = Y_1 + Y_j, \ 2 \leq j \leq n, \]

and the Jacobian of the transformation is 1. Since we are using the location invariant function we can assume that the initial mean is 0. So the likelihoods under the null and the alternative hypotheses are given by
\[ \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} \exp \left[ -\frac{1}{2} \left\{ y_1^2 + \sum_{j=2}^{n} (y_1 + y_j)^2 \right\} \right] dy_1 \] (5.2)

and
\[ \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} \exp \left[ -\frac{1}{2} \left\{ y_1^2 + \sum_{j=2}^{k} (y_1 + y_j)^2 + \sum_{j=k+1}^{n} (y_1 + y_j - \theta)^2 \right\} \right] dy_1 \] (5.3)

respectively. We have to compute the integrals in (5.2) and (5.3) to derive the likelihood ratio. The integrals can be computed easily by completion of squares. Observe
that
\[
\sum_{j=1}^{n} (y_1 + \alpha_j)^2 = n(y_1 + \bar{\alpha})^2 + \sum_{j=1}^{n} (\alpha_j - \bar{\alpha})^2,
\] (5.4)
where $\bar{\alpha}$ is the average of the $\alpha_j$'s. From (5.4) we can see that the integrals reduce to that of a normal density. So (5.2) and (5.3) reduce to
\[
\frac{1}{\sqrt{n(2\pi)^{n-1}/2}} \exp \left[ -\frac{1}{2} \left\{ \sum_{j=2}^{n} (y_j - \bar{y})^2 \right\} \right]
\] (5.5)
and
\[
\frac{1}{\sqrt{n(2\pi)^{n-1}/2}} \exp \left[ -\frac{1}{2} \left\{ \sum_{j=2}^{k} \left( y_j - \bar{y} + \frac{n-k}{n} \theta \right)^2 \sum_{j=k+1}^{n} \left( y_j - \bar{y} - \frac{k}{n} \theta \right)^2 \right\} \right]
\] (5.6)
where $\bar{y} = \sum_{j=2}^{n} y_j / n$. Thus the likelihood ratio is given by
\[
\exp \left[ \delta (kS_n/n - S_k) - \delta^2 k(1 - k/n)/2 \right]
\]
Finally if we use the uniform prior for the location of the change-point we get the statistic $R_k(\delta)$ studied by Pollak and Siegmund.

5.2 Consistency Result

The statistic we derived is used to detect one-sided changes in the mean. Since we are interested in changes of the mean in either direction, we will use the two sided version of this to define our statistic:
\[
S(X_1, \ldots, X_k) = [R_k(\delta) + R_k(-\delta)]/2.
\] (5.7)
We will use the statistic above to define the stopping rules for estimating the number and the locations of the change-points. The following consistency result holds for the stopping rules defined using (5.7). This result follows from the the consistency statement given in Section 4.2.
**Theorem 5.1** Assume that

\[(i) \inf_i \{ \nu_i^{(n)} - \nu_i^{(n)} \} \geq c_n \log n \forall n, \text{ where } c_n \to \infty \text{ as } n \to \infty\]

\[(ii) \inf_i |\theta_i^{n+1} - \theta_i| \geq \frac{\delta}{2} + \epsilon, \text{ for some } \epsilon > 0 \text{ independent of } n.\]

Let \( N_i^{(n)} \) be the stopping rules defined in (4.1) using the statistic defined in (5.7). Let \( \beta_n = c'_n \log n \) for some \( c'_n \) such that \( c'_n \to \infty \) and \( c'_n/c_n \to 0 \) as \( n \to \infty \). Let \( b_n = n^2 \) and let

\[ A_n = \{ \nu_i^{(n)} < N_i^{(n)} < \nu_i^{(n)} + \beta_n, 1 \leq i \leq m_n \& N_{m_n+1}^{(n)} = \infty \}. \]

Then \( P(A_n) \) converges to 1 as \( n \) goes to infinity.

**Proof:** The proof of this result will follow the method given in Section 4.2. Let the setup be as described by [\*] in Section 4.3 with the statistic \( S \) given by (5.7) and the boundary \( b \) as in this theorem. We have to find good upper bounds for \( P_\infty(N \leq n) \) and \( \max_{\beta_n \leq \nu \leq n - \beta_n} P^\nu(N > \nu + \beta_n) \).

From the martingale properties of likelihood ratio we can see that \( R_\infty(\delta) - k \) is a \( P_\infty \) martingale and hence \( S(Y_1, \ldots, Y_k) - k \) is a \( P_\infty \) martingale (Pollak and Siegmund (1991)). Using this property of the statistic we can obtain the following bound

\[ P_\infty(N \leq n) \leq n/b_n. \]

Finally using \( P(B_{in}) \leq P_\infty(N \leq n), b_n = n^2 \) and \( m_n \leq n/\log n \) we get that

\[ \sum_{i=0}^{m_n} P(B_{in}) \leq (m_n + 1)P_\infty(N \leq n) \leq \frac{n}{\log n} \left( \frac{n}{b_n} \right) = \frac{1}{\log n}. \quad (5.8) \]

This gives us bounds for the probability of detecting a change when there isn’t any. Now we need bounds for the probability of not detecting a change when there is a change.

We need \( P^\nu(N > k) \) for \( \beta_n \leq \nu \leq n - \beta_n \) where \( k = \nu + \beta_n \). We can assume without loss of generality that \( \mu > 0 \). Then

\[ P^\nu(N > k) \leq P^\nu(S(Y_1, \ldots, Y_k) < n^2) \]
\[ \begin{align*}
& \leq P_\nu^{(\mu)} \left( R_k(\delta) < 2n^2 \right) \\
& \leq P_\nu^{(\mu)} \left( \exp \left[ \delta \left( \nu \frac{k}{\hat{S}_k} - \hat{S}_k \right) - \delta^2 \nu (1 - \nu/k)/2 \right] < 2n^2 \right) \\
& = P_\nu^{(\mu)} \left( \delta \left( \nu \frac{k}{\hat{S}_k} - \hat{S}_k \right) - \delta^2 \nu (1 - \nu/k)/2 < 2 \log n + \log 2 \right) \\
& = \Phi \left( \frac{2 \log n + \log 2 - \delta (\mu - \delta/2) \nu (1 - \nu/k)}{\sqrt{\delta^2 \nu (1 - \nu/k)}} \right) \\
& \leq \Phi \left( \frac{2 \log n + \log 2 - \delta \epsilon \nu (1 - \nu/k)}{\sqrt{\delta^2 \nu (1 - \nu/k)}} \right). \quad (\star)
\end{align*} \]

The first three inequalities are from the definition of \( N, S(Y_1, \ldots, Y_k) \) and \( R_k(\delta) \) respectively. The first equality is trivial and the second equality follows from normality of \( \nu \frac{k}{\hat{S}_k} - \hat{S}_k \). The final inequality is because \( \mu > \delta/2 + \epsilon \). Since \( \nu \geq \beta_n = c'_n \log n \) and for large \( n \), \( 2 \log n + \log 2 < \delta \epsilon \nu (1 - \nu/k)/2 \), we get that

\[ (\star) \leq \Phi \left( \frac{-\delta \epsilon \nu (1 - \nu/k)/2}{\sqrt{\delta^2 \nu (1 - \nu/k)}} \right) \leq \Phi \left( -\kappa \sqrt{c'_n \log n} \right), \]

where the constant \( \kappa > 0 \) is independent of \( n \). Thus

\[ \sum_{i=1}^{m_n} P(C_i|n) \leq \sum_{i=1}^{m_n} \max_{\beta_n \leq \nu \leq \beta_m} P_\nu^{(\theta_i - \theta_{i-1})}(N > \nu + \beta_n) \leq \frac{n}{\log n} \Phi \left( \kappa \sqrt{c'_n \log n} \right). \quad (5.9) \]

From (5.8) and (5.9) we get that

\[ P(A_k^n) \leq \sum_{i=0}^{m_n} P(B_i|n) + \sum_{i=1}^{m_n} P(C_i|n) \leq \frac{1}{\log n} \left[ 1 + n \Phi \left( \kappa \sqrt{c'_n \log n} \right) \right]. \quad (5.10) \]

It is easy to see that the right hand side of (5.10) converges to zero completing the proof of the theorem. \( \text{Q.E.D.} \)

Although this method is consistent for estimating the number and the locations of the change-points, it requires the knowledge of the minimum absolute change in mean. Also this performs well if the actual absolute change in mean is close to \( \delta \) but might not be as good for other values of change in mean. In order to overcome
these weaknesses we define a different statistic which is a standard normal mixture of $R_k(\delta)$, given by

$$
S(X_1, \ldots, X_k) = \int_{-\infty}^{\infty} R_k(\delta) \exp(-\delta^2/2) d\delta / \sqrt{2\pi} \\
= \sum_{i=0}^{k-1} \exp \left[ \frac{(iS_k/k - S_i)^2}{i(1-i/k) + 1} \right]
$$

(5.11)

The following consistency theorem, which follows the prototype of Statement 4.1, holds for the pseudo-sequential procedure defined with the statistic $S$ in (5.11).

**Theorem 5.2** Assume that

(i) $\inf_i \{\nu_i^{(n)} - \nu_i^{(n)}\} \geq c_n \log n \ \forall n$, where $c_n \to \infty$ as $n \to \infty$

(ii) $\inf_i |\theta_{i+1} - \theta_i| \geq \epsilon$, for some $\epsilon > 0$ independent of $n$.

Let $N_i^{(n)}$ be the stopping rules defined in (4.1) using the statistic defined (5.11). Let $\beta_n = c'_n \log n$ for some $c'_n$ such that $c'_n \to \infty$ and $c'_n/c_n \to 0$ as $n \to \infty$. Let $b_n = n^2$

and let

$$
A_n = \{\nu_i^{(n)} < N_i^{(n)} < \nu_i^{(n)} + \beta_n, 1 \leq i \leq m_n \ & N_{i+1}^{(n)} = \infty\}.
$$

Then $P(A_n)$ converges to 1 as $n$ goes to infinity.

**Proof:** The proof of this theorem is similar to that of Theorem 5.1. We will again use the setup [*] described in Section 4.3, but with the statistic (5.11) and the boundary $b$ given in the theorem. Again we have to find upper bounds for $P_\infty(N \leq n)$ and $\max_{\beta_n \leq \nu \leq n - \beta_n} P^{(\nu)}(N > \nu + \beta_n)$.

In the proof of Theorem 5.1 we said that $R_k(\delta) - k$ is a $P_\infty$ martingale for every $\delta$. Since the statistic $S$ defined by (5.11) is a standard normal mixture of $R_k(\delta)$, the martingale property is preserved, that is, $S(Y_1, \ldots, Y_k) - k$ is a $P_\infty$ martingale. Thus $P_\infty(N \leq n) \leq n/b_n$ and

$$
\sum_{i=0}^{m_n} P(B_{in}) \leq (m_n + 1) P_\infty(N \leq n) \leq \frac{n}{\log n} \left( \frac{n}{b_n} \right) = \frac{1}{\log n}.
$$

(5.12)
This takes care of the first part. For the second part, we need \( P^{(\alpha)}_{\nu}(N > \nu + \beta_n) \) for \( \beta_n \leq \nu \leq n - \beta_n \) which can be obtained as follows.

\[
\begin{align*}
P^{(\alpha)}_{\nu}(N > k) & \leq P^{(\alpha)}_{\nu}(S(Y_1, \ldots, Y_k) < n^2) \\
& \leq P^{(\alpha)}_{\nu} \left( \frac{(\nu \tilde{S}_k/k - \tilde{S}_\nu)^2}{\nu(1 - \nu/k) + 1} < n^2 \right) \\
& \leq P^{(\alpha)}_{\nu} \left( \frac{(\nu \tilde{S}_k/k - \tilde{S}_\nu)^2}{\nu(1 - \nu/k) + 1} < 2 \log n \right) \\
& \leq P^{(\alpha)}_{\nu} \left( \nu \tilde{S}_k/k - \tilde{S}_\nu < \sqrt{2 \log n} \left[ \frac{1}{\nu(1 - \nu/k)} + 1 \right] \right) \\
& \leq \Phi \left( \sqrt{2 \log n} \sqrt{\frac{1}{\nu(1 - \nu/k)} + 1} - \epsilon \sqrt{\nu(1 - \nu/k)} \right) \\
& \leq \Phi \left( -\kappa \sqrt{c_n \log n} \right)
\end{align*}
\]

where \( k = \nu + \beta_n \) and \( \kappa \) a positive constant independent of \( n \). The final inequality follows from the condition that \( \nu \) and \( k - \nu \) are of larger order than \( \log n \). (The reasons here are similar to the arguments that led to (5.9) in Theorem 5.1). So we get

\[
\sum_{i=1}^{m_n} P(C_{i,n}) \leq \sum_{i=1}^{m_n} \max_{\beta_n \leq \nu \leq n - \beta_m} P^{(\theta_i - \theta_{i-1})}_{\nu}(N > \nu + \beta_n) \leq \frac{n}{\log n} \Phi \left( \kappa \sqrt{c_n \log n} \right). \tag{5.13}
\]

The two inequalities (5.12) and (5.13), along with (5.10) and the related arguments in Theorem 5.1, give us the theorem. Q.E.D.

We have proved a consistency statement for the Bayesian procedures both for fixed \( \delta \) and for the mixture. The conditions for the consistency in both the cases are the same except that the fixed \( \delta \) procedure requires the knowledge of a lower bound for the minimal jump in the mean. From these two theorems we can derive various results that follow.
Corollary 5.1 Under the conditions of Theorem 5.1 (or Theorem 5.2) the following results are true for the corresponding estimates.

(i) \( P(\hat{m}_n = m_n) \rightarrow 1 \) as \( n \rightarrow \infty \),

(ii) \( |\hat{\nu}_i^{(n)} - \hat{\nu}_i^{(n)}|/n \) converges to 0 with probability one.

Proof: Observe that \( \hat{m}_n = m_n \) on \( A_n \). Hence (i) follows from the theorem. The other result is trivial from the structure of \( A_n \) and the definition if \( \beta_n \).

Corollary 5.2 Let \( \hat{\mu}_j = (S_{\hat{\nu}_{j+1}} - S_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j) \), \( 0 \leq j \leq \hat{m}_n \), where \( \hat{\nu}_0 = 0 \) and \( \hat{\nu}_{\hat{m}_n+1} = n \). Then, under the conditions of Theorem 5.1 (Theorem 5.2), \( \hat{\mu}_j \) converges to \( \mu_j \) in probability.

Proof: Observe that, \( S_{\hat{\nu}_{j+1}} - S_{\hat{\nu}_j} = (V_{\hat{\nu}_{j+1}} - V_{\hat{\nu}_j}) + (M_{\hat{\nu}_{j+1}} - M_{\hat{\nu}_j}) \), where \( M_k \)'s are the partial sums of the means and \( V_k \)'s are the partial sums of the errors. By Lemma 2.1, \( (V_{k+1} - V_k)/l \) is bounded by \( 1/\sqrt{\log n} \), for all \( l > \log n \) and \( k \), on a set of probability converging to one. By the definition of the set \( A_n \), since \( \hat{\nu}_{j+1} - \hat{\nu}_j \) is greater than \( \log n \) \( (V_{\hat{\nu}_{j+1}} - V_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j) \) converges to zero in probability. By the definition of the set \( A_n \), \( (M_{\hat{\nu}_{j+1}} - M_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j) \) converges to \( \mu_j \) in probability.

5.3 The Unknown Variance Case

In the last two sections we discussed the known variance case. Now we will derive an analogue of \( R_k(\delta) \) for the unknown variance case. Let \( X_1, \ldots, X_n \) be independent normal random variables and let \( S_k = X_1 + \cdots + X_k \). Then the location-scale invariant transformation for this set of random variables is given by

\[
Y_1 = X_1, \quad Y_2 = X_2 - X_1, \quad Y_j = (X_j - X_1)/Y_2, \quad 3 \leq j \leq n.
\]

We want to derive the likelihood ratio statistic, based on \( \{Y_j, 3 \leq j \leq n\} \), for testing the hypothesis that there is no change in mean against the alternative that there is
a change in mean at $k$ and the difference in the means is $\theta$. Since we are using the location-scale invariant statistic we can assume without loss of generality that the initial mean is 0 and the variance is 1. The inverse transformation is given by

$$X_1 = Y_1, \quad X_2 = Y_1 + Y_2, \quad X_j = Y_1 + Y_2Y_j, \quad 3 \leq j \leq n,$$

and the Jacobian of the transformation is $|Y_2|^{n-2}$. So the likelihoods under the two hypotheses are given by

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} |y_2|^{n-2} \exp \left[-\frac{1}{2} \left( y_1^2 + (y_1 + y_2)^2 + \sum_{j=3}^{n} (y_1 + y_2y_j)^2 \right) \right] dy_1 dy_2$$

and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}} |y_2|^{n-2} \exp \left[-\frac{1}{2} \left( y_1^2 + (y_1 + y_2)^2 + \sum_{j=3}^{k} (y_1 + y_2y_j)^2 + \sum_{j=k+1}^{n} (y_1 + y_2y_j - \theta)^2 \right) \right] dy_1 dy_2$$

respectively. Since the integral in (5.15) is more general than that in (5.14) it is enough to evaluate (5.15) and (5.14) would follow.

We will evaluate (5.15) by integrating out $y_1$ first and then $y_2$. Observe that $y_1$ appears only in the exponent. We will complete the squares so that the integral with respect to $y_1$ can be evaluated simply. The exponent can be simplified as follows

$$y_1^2 + (y_1 + y_2)^2 + \sum_{j=3}^{k} (y_1 + y_2y_j)^2 + \sum_{j=k+1}^{n} (y_1 + y_2y_j - \theta)^2$$

$$= \sum_{j=1}^{n} (y_1 + \alpha_j)^2 = n(y_1 + \bar{\alpha})^2 + \sum_{j=1}^{n} (\alpha_j - \bar{\alpha})^2$$

(5.16)

where $\alpha_1 = 0, \alpha_2 = y_2, \alpha_j = y_2y_j, (3 \leq j \leq k), \alpha_j = y_2y_j - \theta (k+1 \leq j \leq n)$ and $\bar{\alpha}$ is the average of the $\alpha_j$'s. From (5.16) we can see that the function of $y_1$ in the integral looks like the normal density. So (5.15) reduces to

$$\int_{-\infty}^{\infty} \frac{|y_2|^{n-2}}{\sqrt{n(2\pi)^{(n-1)/2}}} \exp \left[-\frac{1}{2} \left( \frac{y_2^2 + \sum_{j=3}^{k} (y_2y_j)^2 + \sum_{j=k+1}^{n} (y_2y_j - \theta)^2}{n} \right) \right] \frac{dy_2}{n}$$

(5.17)
We will analyze the exponent in (5.17) as follows.

\[
y_2^2 + \sum_{j=3}^{k} (y_2 y_j)^2 + \sum_{j=k+1}^{n} (y_2 y_j - \theta)^2 - \frac{(y_2 + \sum_{j=3}^{n} y_2 y_j - (n-k)\theta)^2}{n}
\]

\[
= y_2^2 \left\{ 1 + \sum_{j=3}^{n} y_j^2 - \frac{(1 + \sum_{j=3}^{n} y_j)^2}{n} \right\}
\]

\[
-2y_2\theta \left\{ \sum_{j=k+1}^{n} y_j - \frac{(n-k)}{n} \left( 1 + \sum_{j=3}^{n} y_j \right) \right\} + \theta^2 \frac{k(n-k)}{n} \tag{5.18}
\]

Let

\[
V_n^2 = \left\{ 1 + \sum_{j=3}^{n} y_j^2 - \frac{(1 + \sum_{j=3}^{n} y_j)^2}{n} \right\} \quad \text{and} \quad (5.19)
\]

\[
U_{kn} = \left\{ \sum_{j=k+1}^{n} y_j - \frac{(n-k)}{n} \left( 1 + \sum_{j=3}^{n} y_j \right) \right\} \tag{5.20}
\]

Then from (5.18) through (5.20) we can see that the integral (5.17) is given by

\[
\int_{-\infty}^{\infty} \frac{|y_2|^{n-2}}{\sqrt{n(2\pi)^{(n-1)/2}}} \exp \left[ -\frac{1}{2} V_n^2 \left( y_2 - \theta \frac{U_{kn}}{V_n^2} \right)^2 + \frac{1}{2} \theta^2 \left( \frac{U_{kn}^2}{V_n^2} - \frac{k(n-k)}{n} \right) \right] dy_2 \tag{5.21}
\]

\[
= \frac{\exp \left[ \frac{1}{2} \theta^2 \left( \frac{U_{kn}^2}{V_n^2} - \frac{k(n-k)}{n} \right) \right]}{\sqrt{n(2\pi)^{(n-2)/2}}} \times \int_{-\infty}^{\infty} \frac{|y_2|^{n-2}}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} V_n^2 \left( y_2 - \theta \frac{U_{kn}}{V_n^2} \right)^2 \right] dy_2
\]

\[
= \frac{E[|Z + \eta|^{n-2}]}{V_n^{n-1} \sqrt{n(2\pi)^{(n-2)/2}}} \exp \left[ \frac{1}{2} \theta^2 \left( \frac{U_{kn}^2}{V_n^2} - \frac{k(n-k)}{n} \right) \right] \tag{5.22}
\]

where \( Z \) is a standard normal random variable and \( \eta = \theta \frac{U_{kn}}{V_n} \). We can obtain the value of (5.14) by setting \( \theta = 0 \) in (5.22). Thus the likelihood ratio is given by

\[
\frac{E[|Z + \eta|^{n-2}]}{E[|Z|^{n-2}]} \exp \left[ \frac{1}{2} \theta^2 \left( \frac{U_{kn}^2}{V_n^2} - \frac{k(n-k)}{n} \right) \right] \tag{5.23}
\]

where \( Z \) and \( \eta \) are as defined before. Finally the analogue of (5.1) can be derived by using an uniform prior for the location of the change-point and is given by

\[
R_n(\theta) = \sum_{i=0}^{n-1} \frac{E[|Z + \eta|^{n-2}]}{E[|Z|^{n-2}]} \exp \left[ \frac{1}{2} \theta^2 \left( \frac{U_{kn}^2}{V_n^2} - \frac{k(n-k)}{n} \right) \right] \tag{5.24}
\]
Observe that the likelihood in (5.23) doesn't depend on the sign of \( \theta \). This is because \( \{Y_3, \ldots, Y_n\} \) is the same irrespective of the sign of the scaling constant. If you want to differentiate between the two, \( Y_2 \) is defined as \( |X_2 - X_1| \). The likelihood then becomes more complicated and more difficult to evaluate.

Since the statistic in (5.24) does not depend on the sign of \( \theta \) we will use it as the statistic \( S \) to define the stopping rules for the fixed \( \theta \) case. This statistic involves integrals that do not have closed form expressions. So we will try to use the mixture statistic which is evaluated using a standard normal prior for \( \theta \) and taking the expectation of the statistic in (5.24) with respect to it. Since the likelihood under no change doesn't depend on \( \theta \) it is enough to take the mixture of (5.21) and take the ratio to get the statistic. The integral of (5.21) with respect to a standard normal density reduces to

\[
\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{\theta^2}{2} \right) \int_{-\infty}^{\infty} \frac{|y_2|^{n-2}}{\sqrt{n}(2\pi)^{(n-1)/2}} \exp \left[ -\frac{V_n^2 y_2^2}{2} + y_2 \theta U_{k\theta} - \theta^2 \frac{k(n - k)}{2n} \right] dy_2 d\theta
\]

\[
= \int \frac{|y_2|^{n-2}}{\sqrt{n}(2\pi)^{n/2}} \exp \left[ -\frac{1}{2} y_2^2 \left( V_n^2 - \frac{U_{k\theta}^2}{\lambda_{k\theta}^2} \right) \right] \int \exp \left[ -\frac{1}{2} \lambda_{k\theta} \left( \theta - y_2 \frac{U_{k\theta}}{\lambda_{k\theta}} \right)^2 \right] d\theta dy_2,
\]

where \( \lambda_{k\theta} = \sqrt{1 + k(n - k)/n} \). The right hand side in turn reduces to

\[
\int_{-\infty}^{\infty} \frac{\lambda_{k\theta} |y_2|^{n-2}}{\sqrt{n}(2\pi)^{(n-1)/2}} \exp \left[ -\frac{1}{2} y_2^2 \left( V_n^2 - \frac{U_{k\theta}^2}{\lambda_{k\theta}^2} \right) \right] dy_2
\]

\[
= \frac{\lambda_{k\theta} (V_n^2 - \frac{U_{k\theta}^2}{\lambda_{k\theta}^2})^{(n-1)/2}}{\sqrt{n}(2\pi)^{(n-1)/2}} E(|Z|^{n-1}),
\]

where \( Z \) is a standard normal random variable. Thus the mixture statistic is given by

\[
S(X_1, \ldots, X_n) = \sum_{k=0}^{n-1} \frac{V_n^{n-1}}{\lambda_{k\theta} (V_n^2 - \frac{U_{k\theta}^2}{\lambda_{k\theta}^2})^{(n-1)/2}}.
\]

We will now prove a consistency result for the procedure defined with the statistic given in (5.26). The result follows the prototype of Statement 4.1.
CHAPTER 5. BAYESIAN METHOD

Theorem 5.3 Assume that

(i) \( \inf \{ \nu_i^{(n)} - \nu_i^* \} \geq c_n \log n \ \forall n, \ \text{where} \ c_n \to \infty \ \text{as} \ n \to \infty \)

(ii) \( \inf \{ \theta_i^{(n)} - \theta_i \} \geq \frac{\delta}{2} + \epsilon, \ \text{for some} \ \epsilon > 0 \ \text{independent of} \ n. \)

Let \( N_i^{(n)} \) be the stopping rules defined in (4.1) using the statistic defined in (5.26). Let \( \beta_n = c_n' \log n \) for some \( c_n' \) such that \( c_n' \to \infty \) and \( c_n'/c_n \to 0 \) as \( n \to \infty \). Let \( b_n = n^2 \) and let

\[
A_n = \{ \nu_i^{(n)} < N_i^{(n)} < \nu_i^* + \beta_n, 1 \leq i \leq m_n \ \& \ N_{m_n+1}^{(n)} = \infty \}.
\]

Then \( P(A_n) \) converges to 1 as \( n \) goes to infinity.

Proof: We will give a sketch of the proof here. Let the setup \([\ast]\) be as defined in Section 4.3. As in the proof of the previous theorem we can get a bound for the probability of detecting a change when there isn’t any by the martingale property of the likelihood ratio. So it is enough to get a good bound for \( P_\nu^{(u)}(N > \nu + \beta_n) \) for \( \nu > \beta_n \). Observe that \( y_k^2V_n^2 \) is the total sum of squares and \( y_k^2U_k/(k(n-k)/n) \) is the between sum of squares for the two groups \( X_1, \ldots, X_k \) and \( X_{k+1}, \ldots, X_n \). So the statistic \( [V_k^2/\lambda_{ik}(V_k^2 - U_{ik}^2/\lambda_{ik}^2)] \) is a function of the two sample t-statistic \( (t_{ik}) \) given by

\[
\frac{1 + t_{ik}^2/(k-1)}{1 + kt_{ik}/(k-1)(i(k-1)+k)}.
\]

So, if \( i = \nu \) and \( k = \nu + \beta_n \), then

\[
P_\nu^{(u)}(N > \nu + \beta_n) \leq P \left( [V_k^2/\lambda_{ik}(V_k^2 - U_{ik}^2/\lambda_{ik}^2)]^{(k-1)/2} < n^2 \right)
\]

\[
= P \left( (k-1)\log[V_k^2/\lambda_{ik}(V_k^2 - U_{ik}^2/\lambda_{ik}^2)] < 4 \log n \right)
\]

\[
\leq P \left( (k-1)\log[(1 + t_{ik}^2/k)/(1 + t_{ik}^2/(i(k-i)+k))] < 4 \log n \right).
\]

Observe that as \( n \) goes to infinity (and so do \( i \) and \( k-i \)), \( t_{ik}^2/(i(k-i)+k) \) goes to zero. So the probability in the bound above can be bounded by \( P(t_{ik}^2 < 8 \log n) \) which
goes to zero exponentially in $k$. Hence we get that the probability of underestimation goes to zero and the theorem. \hspace{1cm} \textbf{Q.E.D.}

Corollaries 5.1 and 5.2 are true for this procedure too since the conditions of the theorem are the same.
Chapter 6

Likelihood Ratio Method

In the last chapter we considered the pseudo-sequential procedures based on Bayesian arguments. Now we will discuss the likelihood ratio method and derive consistency results for them. These are pseudo-sequential procedures motivated by the CUSUM test used for detecting a change-point. We will be dealing with both the known and the unknown variance case.

6.1 The Known Variance Case

Let $X_1, \ldots, X_n$ be the data and let $S_k = X_1 + \cdots + X_k$. Then the likelihood ratio statistic for testing the hypothesis that there is no change in $X_i, \ldots, X_j$ against the alternative that there is change at $i$ and the difference in mean is $\delta$ is given by

$$Z_{ij}(\delta) = \delta[(iS_j/j - S_i) - (\delta/2)i(1 - i/j)].$$

So for testing the hypothesis that there is no change-point in $X_1, \ldots, X_k$ against the alternative that there is change of magnitude $\delta > 0$ we use the statistic $S$ given below

$$S(X_1, \ldots, X_k) = \max_{0 < i < j < k} \max \{Z_{ij}(\delta), Z_{ij}(-\delta)\} \tag{6.1}$$
CHAPTER 6. LIKELIHOOD RATIO METHOD

We can use the statistic $S$ in (6.1) to define the stopping rules defined in (4.1) and estimate the number and locations of the change-points as in (4.2). The following consistency result holds for the procedure with the above $S$.

**Theorem 6.1** Assume that

\[(i) \inf_i \left\{ \nu^{(n)}_{i+1} - \nu^{(n)}_i \right\} \geq c_n \log n \quad \forall n, \text{ where } c_n \to \infty \text{ as } n \to \infty\]

\[(ii) \inf_i |\theta_{i+1} - \theta_i| \geq \frac{\delta}{2} + \epsilon, \text{ for some } \epsilon > 0 \text{ independent of } n.\]

Let $N_i^{(n)}$ be the stopping rules defined in (4.1) using the statistic defined in (6.1). Let $\beta_n = c'_n \log n$ for some $c'_n$ such that $c'_n \to \infty$ and $c'_n/c_n \to 0$ as $n \to \infty$. Let $b_n = 3 \log n$ and let

$$
A_n = \left\{ \nu^{(n)}_i < N_i^{(n)} < \nu^{(n)}_i + \beta_n, 1 \leq i \leq m_n \& N_{m_n+1}^{(n)} = \infty \right\}.
$$

Then $P(A_n)$ converges to 1 as $n$ goes to infinity.

**Proof:** We will use [*] described Section 4.3, with the statistic $S$ defined in (6.1) and boundary $b$ as in the theorem, and the method of proof in Section 4.2.

Let $\tilde{Z}_{ij}$ be defined the same way as $Z_{ij}$, with the partial sums $\tilde{S}_i$'s instead of $S_i$'s. We will bound the probability of detecting a change-point when there isn't any as follows.

$$
P_\infty(N \leq n) \leq \sum_{1 \leq i < j \leq n} P_\infty \left( \tilde{Z}_{ij}(\delta) > b_n \right)
$$

$$
= \sum_{1 \leq i < j \leq n} \left[ 1 - \Phi \left( \frac{b_n + \delta(1 - i/j)2}{\sqrt{\delta^2 i(1 - i/j)}} \right) \right]
$$

$$
\leq \sum_{1 \leq i < j \leq n} \left[ 1 - \Phi \left( \sqrt{2b_n} \right) \right].
$$

The first equality follows from bounding the probability of the union by the sum of the probabilities. The equality is from normality of the random variable involved.
The second inequality is by minimizing the argument over all \( i, j \). Since \( b_n = 3 \log n \) we get that

\[
nP_\infty(N \leq n) \to 0 \text{ as } n \to \infty.
\] (6.2)

Now to bound the probability of missing a change-point, we need \( P^{(\mu)}_\nu(N > k) \) for \( \beta_n \leq \nu \leq n - \beta_n \) where \( k = \nu + \beta_n \). We can assume without loss of generality that \( \mu > 0 \).

\[
P^{(\mu)}_\nu(N > k) \leq P_\nu(S(Y_1, \ldots, Y_k) < 3 \log n)
= P^{(\mu)}_\nu\left(\max_{1 \leq i < j \leq k} \tilde{Z}_{ij}(\delta) < 3 \log n\right)
\leq P^{(\mu)}_\nu\left(\tilde{Z}_{\nu k}(\delta) < 3 \log n\right)
\leq P^{(\mu)}_\nu\left(\delta(\nu \tilde{S}_k/k - \tilde{S}_\nu) - \delta^2 \nu(1 - \nu/k)/2 < 3 \log n\right)
= \Phi\left(\frac{3 \log n - \delta(\mu - \delta/2)\nu(1 - \nu/k)}{\sqrt{\delta^2 \nu(1 - \nu/k)}}\right)
\leq \Phi\left(\frac{3 \log n - \delta \nu(1 - \nu/k)}{\sqrt{\delta^2 \nu(1 - \nu/k)}}\right). \quad (*)
\]

For large \( n \) observe that \( 3 \log n \leq \delta \nu(1 - \nu/k)/2 \). Hence for large \( n \)

\[
(*) \leq \Phi \left(-\kappa \sqrt{c_n \log n}\right)
\] (6.3)

for some positive constant \( \kappa \) independent of \( n \). From (6.2) and (6.3) we can get the proof of this theorem by arguments similar to that in Theorem 5.1. \( \text{Q.E.D.} \)

This procedure has the same weaknesses as the Bayesian procedure with fixed \( \delta \), that is, we need to know a lower bound for the minimal jump in means and the procedure doesn’t perform well for jumps in mean different from the hypothesised value. So we will define a new statistic which uses the generalized likelihood ratio statistic to test the hypothesis of no change until \( j \) against a change at \( i \) given by

\[
Z_{ij} = (iS_j/j - S_i)^2/[i(1 - i/j)]
\]
The statistic $S$ in this case is given by

$$S(X_1, \ldots, X_k) = \max_{0 < i < j < k} Z_{ij}$$

(6.4)

We get the following theorem for this procedure.

**Theorem 6.2** Assume that

(i) $\inf_i \left\{ \nu_{i+1}^{(n)} - \nu_i^{(n)} \right\} \geq c_n \log n \ \forall n$, where $c_n \to \infty$ as $n \to \infty$

(ii) $\inf_i |\theta_{i+1} - \theta_i| \geq \epsilon$, for some $\epsilon > 0$ independent of $n$.

Let $N_i^{(n)}$ be the stopping rules defined in (4.1) using the statistic defined in (6.4). Let $\beta_n = c'_n \log n$ for some $c'_n$ such that $c'_n \to \infty$ and $c'_n/c_n \to 0$ as $n \to \infty$. Let $b_n = 6 \log n$ and let

$$A_n = \left\{ \nu_i^{(n)} < N_i^{(n)} < \nu_i^{(n)} + \beta_n, 1 \leq i \leq m_n \ & N_{m_n+1}^{(n)} = \infty \right\}.$$

Then $P(A_n)$ converges to 1 as $n$ goes to infinity.

**Proof:** The proof of this theorem by mimicking the arguments in the proof of Theorem 6.1. We will use the setup [•] with the statistic $S$ in (6.4) and $b$ as in theorem.

Let $\tilde{Z}_{ij}$ be defined with $S$’s replaced by $\tilde{S}$’s in the definition of $Z_{ij}$. By bounding the probability of the union by the sum of the probabilities we get that

$$P_\infty(N \leq n) \leq \sum_{1 \leq i < j \leq n} P_\infty(\tilde{Z}_{ij} > b_n) = 2 \sum_{1 \leq i < j \leq n} \left[ 1 - \Phi \left( b_n^{1/2} \right) \right].$$

The inequality follows from bounding the probability of the union by the sum of the probabilities. Since $\tilde{Z}_{ij}$ is the square of a standard normal random variable the equality follows. Since $b_n = 6 \log n$ we get that

$$nP_\infty(N \leq n) \to 0 \text{ as } n \to \infty.$$  

(6.5)
Now for the second part, that is, to find an upper bound for the probability of missing a change-point, we need \( P^{(\mu)}_\nu(N > \nu + \beta_n) \) for \( \beta_n \leq \nu \leq n - \beta_n \). We can assume without loss of generality that \( \mu > 0 \). By following the arguments of the previous theorem, we get,

\[
P^{(\mu)}_\nu(N > k) \leq P_\nu(S(Y_1, \ldots, Y_k) < 6 \log n)
\leq P^{(\mu)}_\nu(\tilde{Z}_{nk} < 6 \log n)
\leq P^{(\mu)}_\nu((\nu \tilde{S}/k - \tilde{S}_\nu)/\sqrt{\nu(1 - \nu/k)} < \sqrt{6 \log n})
= \Phi \left( \sqrt{6 \log n - \frac{\mu \nu(1 - \nu/k)}{\nu(1 - \nu/k)}} \right)
\leq \Phi \left( \sqrt{6 \log n - \epsilon \sqrt{\nu(1 - \nu/k)}} \right). \tag{*}
\]

For large \( n \) observe that \( 3 \log n \leq \delta \epsilon \nu(1 - \nu/k)/2 \). Hence for large \( n \)

\[
(*) \leq \Phi \left( -\kappa \sqrt{c_n \log n} \right) \tag{6.6}
\]

for some positive constant \( \kappa \) independent of \( n \). From (5.5) and (6.6) we can get the proof of this theorem by arguments similar to that in Theorem 6.1. \( \text{Q.E.D.} \)

As in the previous chapter we get the following corollaries. Since the conditions in both these theorems are the same as the ones in Theorem 5.1, the proof are the same.

**Corollary 6.1** Under the conditions of Theorem 6.1 (or Theorem 6.2) the following results are true for the corresponding estimates.

(i) \( P(\tilde{m}_n = m_n) \to 1 \) as \( n \to \infty \),

(ii) \( |\tilde{\nu}_i^{(n)} - \nu_i^{(n)}|/n \) converges to 0 with probability one.

**Corollary 6.2** Let \( \hat{\mu}_j = (S_{\tilde{\nu}+1} - S_{\tilde{\nu}})/(\tilde{\nu}_{j+1} - \tilde{\nu}_j), 0 \leq j \leq \tilde{m}_n \), where \( \tilde{\nu}_0 = 0 \) and \( \tilde{\nu}_{\tilde{m}_n+1} = n \). Then, under the conditions of Theorem 6.1 (Theorem 6.2), \( \hat{\mu}_j \) converges to \( \mu_j \) in probability.
6.2 The Unknown Variance Case

We will now consider the unknown variance case. We will derive the consistency result for the pseudo-stopping rules defined using the generalized likelihood ratio statistic and the results for the fixed alternative can be proved similarly.

We want to test the hypothesis that \( X_1, \ldots, X_j \) have the same mean against the alternative that the \( X_1, \ldots, X_i \) and \( X_{i+1}, \ldots, X_j \) have different means. The generalized likelihood ratio statistic, which is the same as the two sample t statistic, for this problem is given by

\[
Z_{ij} = \frac{|iS_j/j - S_i|/\hat{\tau}_{ij}[i(1 - i/j)]^{1/2},}
\]

where \( S_i = X_1 + \cdots + X_i \) and

\[
\hat{\tau}_{ij}^2 = \left[ \sum_{l=1}^{i} (X_l - S_l/i)^2 + \sum_{l=i+1}^{j} (X_l - (S_j - S_l)/(j-i))^2 \right] / (j-2).
\]

For testing no change against change we use the maximum of \( Z_{ij} \) over \( i \). We will use the statistic \( S \) to define the stopping rules given by (4.1), where

\[
S(X_1, \ldots, X_k) = \max_{0 < i < k} Z_{ik}. \tag{6.7}
\]

We can prove the following consistency statement for the stopping rules defined using the statistic \( S \) in (6.7)

**Theorem 6.3** Assume that

(i) \( \inf_i \{ \nu_i^{(n)} - \nu_i^{(n)} \} \geq c_n n^{2/3} \forall n, \text{ where } c_n \to \infty \text{ as } n \to \infty \)

(ii) \( \inf_i \ |\theta_{i+1} - \theta_i| \geq \epsilon, \text{ for some } \epsilon > 0 \text{ independent of } n. \)

Let \( N_i^{(n)} \) be the stopping rules defined in (4.1) using the statistic defined in (6.7). Let \( \beta_n = c' n^{2/3} \) for some \( c' \) such that \( c' \to \infty \) and \( c'/c_n \to 0 \) as \( n \to \infty \). Let \( b_n = n^{1/3} \) and let

\[
A_n = \left\{ \nu_i^{(n)} < N_i^{(n)} < \nu_i^{(n)} + \beta_n, 1 \leq i \leq m_n \& N_{m_n+1}^{(n)} = \infty \right\}.
\]

Then \( P(A_n) \) converges to 1 as \( n \) goes to infinity.
Proof: The proof follows the structure given in Section 4.2. We will use the setup [*] with $S$ defined by (6.7) and $b$ as in the theorem. In order to find an upper bound for the probability of detecting a change-point when there isn’t any, we need the following lemma.

Lemma 6.1 Let $U_1, \ldots, U_n$ be i.i.d. standard normal random variables. Let $t_{ij}$ be the two sample t-statistic to test for the equality of the means of the two groups \{$U_1, \ldots, U_i$\} and \{$U_{i+1}, \ldots, U_j$\}, where $t_{12} = 0$. Then for all $b > \alpha \sqrt{\log n}$, $\alpha > 0$ independent of $n$,

$$\frac{c_1}{b} \leq P \left( \max_{1 \leq i < j \leq n} t_{ij} > b \right) \leq \frac{c_2}{b}$$

where $c_1$ and $c_2$ are constants independent of $n$.

Proof: Observe that

$$P(t_{13} > b) \leq P \left( \max_{1 \leq i < j \leq n} t_{ij} > b \right) \leq \sum_{1 \leq i < j \leq n} P(t_{ij} > b) \quad (6.8)$$

Let $W_i$ have a t-distribution with $i$ degrees of freedom. Then from (6.8) we get that

$P(W_1 > b)$ and $\sum_{i=1}^{n-2} (i + 1)P(W_i > b)$ give us a lower and upper lower bound for the probability.

Since a t-distribution with one degree of freedom is a Cauchy distribution, we get that $P(W_1 > b) = (\pi/2 - \arctan b)/\pi$ goes to zero at rate $1/b$. Thus we get the lower bound.

We will now get the upper bound. Let $B(k, l)$ be the beta function given. For $i \geq 2$

$$P(W_i > b) = \int_b^\infty (\sqrt{i}B(1/2, i/2))^{-1}(1 + t^2/i)^{-i+1} \, dt$$

$$= \int_{b/\sqrt{i}}^\infty (B(1/2, i/2))^{-1}(1 + tr)^{-i+1} \, dt$$

$$\leq \frac{\sqrt{i}}{b} \int_{b/\sqrt{i}}^\infty (B(1/2, i/2))^{-1}t(1 + t^2)^{-i+1} \, dt$$

$$= \frac{i}{i - 1} (\sqrt{i}B(1/2, i/2))^{-1}b^{-1}(1 + b^2/i)^{-i+1}.$$
Since \( \sqrt{i}B(1/2, i/2) \) converges to \( \sqrt{2\pi} \) we get that \( P(W_i > b) \) is bounded by \( \kappa/[b(1 + b^2/i)^{\frac{i+1}{2}}] \) for some constant \( \kappa \) independent of \( n \) and \( b \).

We will now prove that, for \( b > 3\sqrt{\log n} \), \( \sum_{i=2}^{n-2} (i+1)P(W_i > b) \) goes to zero at a faster rate than \( 1/b \). We will do it by dividing the sum into three factors as follows

\[
\sum_{i=2}^{n-2} (i+1)P(W_i > b) = \frac{\kappa}{b} \sum_{i=2}^{n-2} \frac{(i+1)}{(1 + b^2/i)^{\frac{i+1}{2}}}
= \frac{\kappa}{b} \left[ \sum_{i=2}^{9} \frac{(i+1)}{(1 + b^2/i)^{\frac{i+1}{2}}} + \sum_{i=10}^{b^4} \frac{(i+1)}{(1 + b^2/i)^{\frac{i+1}{2}}} + \sum_{i=b^4}^{n-2} \frac{(i+1)}{(1 + b^2/i)^{\frac{i+1}{2}}} \right].
\]

We will show the term in the square brackets converges to zero.

We will first deal with the easy one which is the final term.

\[
\sum_{b^4}^{n-2} \frac{(i+1)}{(1 + b^2/i)^{\frac{i+1}{2}}} = \sum_{b^4}^{n-2} (i+1) \exp \left\{ -\frac{i-1}{2} \log(1 + b^2/i) \right\}
\leq \sum_{b^4}^{n-2} \frac{(i+1)}{2} \left\{ \frac{i-1}{i} b^2 - \frac{i-1}{i^2} b^4 \right\}
\leq n^2 \exp\{1 - b^2\}.
\]

The first equality from the bound for \( P(W_i > b) \) derived earlier. The equality is trivial. The second inequality is using \( \log(1 + x) \geq x - x^2/2 \) and the final inequality follows from \( (k - 1)b^4/k^2 \leq 1 \). Since \( b > 3 \log n \) the bound for this sum goes to zero.

Now we will deal with the second term. Observe that

\[
(1 + b^2/i)^{\frac{i+1}{2}} \geq \left[ \left( \frac{i-1}{9} \right) (b^2/i)^{9} \right]^{1/2} \geq \frac{b^9}{10^5}.
\]

So

\[
\sum_{i=10}^{b^4} \frac{(i+1)}{(1 + b^2/i)^{\frac{i+1}{2}}} \leq \sum_{i=10}^{b^4} (i+1)10^5/b^9 \leq 10^5/b.
\]

Clearly this bound converges to zero.
Finally, for $2 \leq i \leq 9$, $(1 + b^2 / i)^{\frac{i+1}{2}} \geq (b / \sqrt{i})^{i-1} \geq b / 3^8$. So the bound for the first term is
\[
\sum_{i=2}^{9} \frac{(i+1)}{(1 + b^2 / i)^{\frac{i+1}{2}}} \leq 36 \times 3^8 / b
\]
which clearly converges to zero.

From these we get that $\sum_{i=2}^{n-2} (i+1)P(W_i > b)$ goes to zero at a faster rate than $1/b$. Since the upper bound is $\sum_{i=1}^{n-2} (i+1)P(W_i > b)$ and $P(W_1 > b)$ goes to zero at rate $1/b$, we get the lemma. \(\square\)

From this lemma we can prove that $P_{\infty}(N \leq n) \leq \kappa / b_n$ where $b_n$ is the boundary in the theorem and $\kappa > 0$ is a constant independent of $n$ and $b_n$. So
\[
\sum_{i=0}^{m_n} P(B_{i,n}) \leq (m_n + 1)P_{\infty}(N \leq n) \leq \frac{n}{\beta_n} \left( \frac{\kappa}{b_n} \right) = \frac{1}{c_n}. \tag{6.9}
\]
This bounds the probability of detecting a change-point when there isn’t any.

We have to find a bound for the probability of missing a change-point. Let $k = \nu + \beta_n$. Let $\mu$ be the difference in the means. Let $U$ and $V$ be independent random variables, where $U$ is a $N(\mu',1)$, with $\mu' = \mu / \sqrt{1 / \nu + 1 / \beta_n}$, and $(k-2)V^2$ is a chi-squared random variable with $k-2$ degrees of freedom. Then
\[
P_{\nu}^{(\mu)}(N > \nu + \beta_n) \leq P_{\nu}^{(\mu)}(Z_{\nu k} \leq b_n) = P(|U|/V < b_n) \leq P(|U| \leq 2b_n) + P(V > 2) \leq \Phi(2b_n - \delta \beta_n^{1/2}) + \exp\{-{(k-2)/3}\},
\]
where the final inequality comes from computing the probability of a chi-squared random variable larger than a constant ($> 0$) times its degree of freedom. We get that $b_n$ should increase at slower rate than $\beta_n^{1/2}$. From (6.9) we can see that $b_n \beta_n$ should increase at a faster rate than $n$. So the best choice of $b_n$ and $\beta_n$ are as specified.
For this choice we get
\[
\sum_{i=1}^{m_n} P(C_{i,n}) \leq \sum_{i=1}^{m_n} \max_{\beta_n \leq \nu \leq \beta_m} P_{\nu}^{(\theta_i - \theta_{i-1})} (N > \nu + \beta_n) \leq n^{1/3} (\Phi(\beta_n) + \exp(-n^{1/3})).
\]  
(6.10)

Since the bounds given by (6.9) and (6.10) converge to zero, we get that \( P(\mathcal{A}_n^c) \) converges to zero. Hence the consistency result follows. Q.E.D.

From this consistency theorem we get the following corollaries, which are consistency statements for the estimates of the number and the locations of the change-points as well as the estimates of the means.

**Corollary 6.3** Under the conditions of Theorem 6.3 the following results are true for the corresponding estimates.

(i) \( P(\hat{m}_n = m_n) \to 1 \) as \( n \to \infty \),

(ii) \( |\hat{\nu}_i^{(n)} - \nu_i^{(n)}|/n \) converges to 0 with probability one.

**Corollary 6.4** Let \( \hat{\mu}_j = (S_{\hat{\nu}_{j+1}} - S_{\hat{\nu}_j})/(\hat{\nu}_{j+1} - \hat{\nu}_j), \) \( 0 \leq j \leq \hat{m}_n \), where \( \hat{\nu}_0 = 0 \) and \( \hat{\nu}_{\hat{m}_n+1} = n \). Then, under the conditions of Theorem 6.1 (Theorem 6.2), \( \hat{\mu}_j \) converges to \( \mu_j \) in probability.

This procedure estimates the variance after every time the procedure stops. Since the variance remains a constant, this results in a loss of efficiency in the sense that it takes the procedure longer to detect the next change-point and requires the minimal distance between the change-points. We can make the procedure more efficient by using a good estimate of the variance before starting the detection of the change-points. One such estimate would be an estimate from a pilot study. One can also use the initial part of the data as a training sample and use the variance estimate from it throughout the procedure. We can also update the estimate as we look at more data. These procedures would be consistent with almost the same conditions as the known variance case.
Chapter 7

Numerical Results

We proved consistency results for the procedures we discussed in the earlier chapters. These results compare the large sample properties of the different procedures. We are also interested in the small sample performance of the procedures. We will now compare the performance of the different procedures described so far, using Monte Carlo simulations. We will also try some ad hoc modifications of some of the procedures aimed at improving their efficiency. We will conduct a factorial experiment to compare the effects of different factors that are involved in the problem, such as, the number of change-points and the size of the jump in the mean. Finally we will consider the example of the Nile data (Cobb (1978)).

7.1 Monte Carlo Results

We proved consistency results for four types of procedures and now we want to compare the performances of the procedures we consider. The pseudo-sequential procedures we studied, have a fixed $\delta$ as well as the generalized version.

Since three of the procedures use tests to detect change-points the issue of choosing the critical values for the test is important. There are two ways of choosing the critical
values. The first is the fixed boundary case (FB), where a universal level $\alpha$ critical value is chosen for the whole procedure, that is, we will choose $b_n$ (data size is $n$) such that the probability of detecting a change in the data, when there isn't any is at most $\alpha$. The second is the variable boundary case (VB), where we will update the boundary depending on the size of the subset we are dealing with, that is, $b$ is chosen according to the segment size for the binary segmentation procedure and the $n - N_j$, where $N_j$ is the last time when the procedure stopped, for the pseudo-sequential procedure. For choosing the critical values, we will use all probability approximations which are discussed in Chapter A of the appendix. We will use the variable critical levels in our simulations. For the procedure based on the Schwarz criterion, one can set the cost function $\alpha_n$ such that a specified error rate can be achieved. But since the error probability is not a simple function of the cost we will not do it. We will set the cost to be identically 1 in our simulation.

We will consider two modifications of the mixture Bayesian method given as follows. The stopping rules say that between two consecutive values $N_{j-1}$ and $N_j$, there is a change-point. So, instead of estimating $\nu_j$ by $N_j$, we can estimate the location by the maximum likelihood estimate. Let $\hat{\nu}_j$ be the estimate. Then we can either proceed from $N_j + 1$ onwards or use the data between $\hat{\nu}_j + 1$ and $N_j$ as a training sample for detecting the next change-point. We can do the same for the other pseudo-sequential procedures too, but since the results are expected to be similar we will not do it.

We will run the simulations as follows. For every mean function, we will generate 1000 random samples with that mean function. We will use all the procedures to estimate the number of change-points and their locations. For all the procedures and mean functions we will get the mean absolute deviation of the estimated number of change-points from the true number. These two help us compare the performance of the procedures as far as the number change-points is concerned. We are also interested in comparing the performance of the procedures in estimating the locations
of the change-points as well as the mean function. There are two statistics we use for this. The first one is the integrated mean squared error (IMSE), \(i.e.\) it is the mean of \((\mu - \hat{\mu})^2\). The smaller the IMSE the better the procedure is in estimating the mean function. The second one is the penalized mean absolute deviation of the estimated locations (PMAD). This is helpful in comparing the procedures in terms of estimating the locations of the change-points and is defined as follows. We penalize underestimation or overestimation in calculating the mean absolute deviation. Let \(m\) be the true number of change-points and \(\hat{n}\) the estimated number of change-points. Let \(\nu_i\)'s and \(\hat{\nu}_i\)'s denote the true and estimated locations of the change-points. Let \(d_i\)'s be the distances between consecutive change-points. If the \(\hat{n}\) is less than \(m\) then PMAD is defined as

\[
m^{-1} \left[ \sum_{j=1}^{\hat{n}} \inf_{i} |\hat{\nu}_j - \nu_i| + \sum_{l=1}^{m-\hat{n}} d(l) \right],
\]

where the infimum is taken over the eligible set (because in the pseudo-sequential case the location estimate cannot be that of a change-point in the data not observed). Since it is most likely that the procedure misses a changed segment with the least amount of data in it we penalize accordingly. If the \(\hat{n}\) is at least as large as \(m\) then PMAD is defined as

\[
m^{-1} \left[ \sum_{i=1}^{m} \inf_{j} |\nu_i - \hat{\nu}_j| + \sum_{l=1}^{m-m-n} d(m+1-l) \right],
\]

where the infimum is taken over the eligible set (same reason as before). Since it is most likely that the procedure detects a change-point in the longest segment we penalize accordingly.

The different procedures and abbreviations we will use for the simulations are given in Table 7.1. There are six mean functions \((f_1 - f_6)\) that we will use for the simulations. Since it is infeasible to do a large scale study for the procedure based on the Schwarz criterion we will run it only for the mean functions \(f_1 - f_3\). There are three change-points in \(f_1 - f_3\) at 25, 50 and 75. The data size is 100. The means are


CHAPTER 7. NUMERICAL RESULTS

\[
\begin{array}{|l|}
\hline
\text{Procedure} & \text{Notation} \\
\hline
\text{Binary Segmentation Procedure} & BS \\
\text{Procedure based on the Schwarz Criterion} & SC \\
\text{Bayesian Method - Mixture} & BM \\
\text{Generalized Likelihood Ratio Method} & LR \\
\text{Bayesian Method - } \delta = 1 & BM - \delta \\
\text{Bayesian Method - Modification 1} & BM - M1 \\
\text{Bayesian Method - Modification 2} & BM - M2 \\
\hline
\end{array}
\]

Table 7.1: Procedures and Notations.

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Range} & 1 - 25 & 26 - 50 & 51 - 75 & 76 - 100 \\
\hline
\mu_i & 0.5 & -0.5 & 0.5 & -0.5 \\
\mu_i & 1 & 2 & 3 & 4 \\
\mu_i & 0.75 & -0.75 & 0.75 & -0.75 \\
\hline
\end{array}
\]

Table 7.2: (a) Mean functions $f_1$, $f_2$ and $f_3$.

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Range} & 1 - 40 & 41 - 80 & 81 - 120 & 121 - 160 & 161 - 200 & 201 - 240 \\
\hline
\mu_i & 0.5 & -0.5 & 0.5 & -0.5 & 0.5 & 0.5 \\
\mu_i & 1 & -1 & 1 & -1 & 1 & -1 \\
\hline
\end{array}
\]

Table 7.2: (b) Mean functions $f_4$ and $f_5$.

\[
\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
\text{Range} & 1 - 20 & 21 - 40 & 41 - 60 & 61 - 80 & 81 - 100 & 101 - 120 \\
\hline
\mu_i & 1 & -1 & 1 & -1 & 1 & -1 \\
\hline
\end{array}
\]

Table 7.2: (c) Mean function $f_6$. 
CHAPTER 7. NUMERICAL RESULTS

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Estimated number of change-points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>BS</td>
<td>268</td>
</tr>
<tr>
<td>SC</td>
<td>174</td>
</tr>
<tr>
<td>BM</td>
<td>208</td>
</tr>
<tr>
<td>LR</td>
<td>244</td>
</tr>
<tr>
<td>BM - δ</td>
<td>126</td>
</tr>
<tr>
<td>BM - M2</td>
<td>208</td>
</tr>
</tbody>
</table>

Table 7.3: Frequency distribution of the estimated number of change-points for the mean function $f_1$.

given in Table 7.2 (a). The mean functions $f_4$ and $f_5$ have five change-points at 40, 80, 120, 160 and 200. The data size is 240 and the means are tabulated in Table 7.2 (b). Finally the mean function $f_6$ has ten change-points at 20, 40, 60, 80, 100, 120, 140, 160, 180 and 200. The data size is 220. The means are given in Table 7.2 (c).

The frequency distribution of the estimated number of change-points for the different procedures are given in Tables 7.3 through 7.8 and correspond to the mean functions $f_1$ through $f_6$. Table 7.9 gives the mean absolute deviation of the estimated number of change-points from the true number. The histograms of the integrated mean squared error and the mean absolute deviations are given in Figures 7.1 through 7.7. In all these tables the frequency distribution of the number of change-points detected for the procedure BM - M1 is same as that of BM. The difference in the two procedures is in IMSE and PMAD because the estimated locations for the two procedures are different.
CHAPTER 7. NUMERICAL RESULTS

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Estimated number of change-points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>BS</td>
<td>57</td>
</tr>
<tr>
<td>SC</td>
<td>50</td>
</tr>
<tr>
<td>BM</td>
<td>172</td>
</tr>
<tr>
<td>LR</td>
<td>177</td>
</tr>
<tr>
<td>BM - δ</td>
<td>91</td>
</tr>
<tr>
<td>BM - M2</td>
<td>57</td>
</tr>
</tbody>
</table>

Table 7.4: Frequency distribution of the estimated number of change-points for the mean function $f_2$.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Estimated number of change-points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>BS</td>
<td>11</td>
</tr>
<tr>
<td>SC</td>
<td>3</td>
</tr>
<tr>
<td>BM</td>
<td>2</td>
</tr>
<tr>
<td>LR</td>
<td>3</td>
</tr>
<tr>
<td>BM - δ</td>
<td>1</td>
</tr>
<tr>
<td>BM - M2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 7.5: Frequency distribution of the estimated number of change-points for the mean function $f_3$.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Estimated number of change-points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>BS</td>
<td>106</td>
</tr>
<tr>
<td>BM</td>
<td>16</td>
</tr>
<tr>
<td>LR</td>
<td>36</td>
</tr>
<tr>
<td>BM - δ</td>
<td>11</td>
</tr>
<tr>
<td>BM - M2</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 7.6: Frequency distribution of the estimated number of change-points for the mean function $f_4$. 
Table 7.7: Frequency distribution of the estimated number of change-points for the mean function $f_5$. 

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Estimated number of change-points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td>BS</td>
<td>612</td>
</tr>
<tr>
<td>BM</td>
<td>985</td>
</tr>
<tr>
<td>LR</td>
<td>914</td>
</tr>
<tr>
<td>BM - $\delta$</td>
<td>974</td>
</tr>
<tr>
<td>BM - M2</td>
<td>932</td>
</tr>
</tbody>
</table>

Table 7.8: (a) Frequency distribution of the estimated number of change-points for the mean function $f_5$. 

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Estimated number of change-points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 - 4</td>
</tr>
<tr>
<td>BS</td>
<td>80</td>
</tr>
<tr>
<td>BM</td>
<td>0</td>
</tr>
<tr>
<td>LR</td>
<td>0</td>
</tr>
<tr>
<td>BM - $\delta$</td>
<td>0</td>
</tr>
<tr>
<td>BM - M2</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.8: (b) Frequencies in the range 0 - 4 for the binary segmentation procedure.
### Table 7.9: Mean absolute deviations of the estimated number of change-points from the true number.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
<th>$f_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS</td>
<td>1.828</td>
<td>0.636</td>
<td>0.564</td>
<td>2.514</td>
<td>0.502</td>
<td>1.610</td>
</tr>
<tr>
<td>SC</td>
<td>1.147</td>
<td>0.775</td>
<td>0.199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM</td>
<td>1.787</td>
<td>1.113</td>
<td>0.365</td>
<td>1.878</td>
<td>0.015</td>
<td>0.961</td>
</tr>
<tr>
<td>LR</td>
<td>1.750</td>
<td>1.056</td>
<td>0.304</td>
<td>2.106</td>
<td>0.089</td>
<td>0.469</td>
</tr>
<tr>
<td>BM - $\delta$</td>
<td>1.467</td>
<td>0.954</td>
<td>0.189</td>
<td>1.461</td>
<td>0.026</td>
<td>0.803</td>
</tr>
<tr>
<td>BM - M2</td>
<td>1.332</td>
<td>0.832</td>
<td>0.115</td>
<td>0.822</td>
<td>0.077</td>
<td>0.131</td>
</tr>
</tbody>
</table>

From the frequency distributions of the estimated number of change-points we can infer the following. The pseudo-sequential procedures generally underestimate the number of change-points if the data is not sufficient to distinguish between the different means. If the data size is sufficient they are very efficient in detecting the changes. The over estimation can be explained by the error rate allowed for detecting a change when there isn’t any. The second modification to the Bayesian method which uses the data between the estimated location and the time at which the procedure stopped is very efficient in detecting the change-points. We expect similar improvement for modifications of the other pseudo-sequential procedures. The binary segmentation procedure has both an underestimation as well as an overestimation problem. Though we don’t have enough data for the procedure based on the Schwarz criterion, we can still see that it has a similar problem. The problem of underestimation is not as severe as that of the other procedures. The problem of overestimation for the procedure SC can be over come by choosing the cost function $\alpha_n$ suitably so that the probability of detecting a change-point when there isn’t any can be controlled. In the following pages we have the histograms of the IMSE and PMAD for different procedures.
Figure 7.1: (a) The histograms of IMSE and PMAD for the mean functions $f_1 - f_3$ for the binary segmentation procedure.
Figure 7.1: (b) The histograms of IMSE and PMAD for the mean functions $f_4 - f_6$ for the binary segmentation procedure.
Figure 7.2: The histograms of IMSE and PMAD for the mean functions $f_1 - f_3$ for the procedure based on the Schwarz criterion.
Figure 7.3: (a) The histograms of IMSE and PMAD for the mean functions $f_1 - f_3$ for the mixture Bayesian procedure.
Figure 7.3: (b) The histograms of IMSE and PMAD for the mean functions $f_4 - f_6$ for the mixture Bayesian procedure.
Figure 7.4: (a) The histograms of IMSE and PMAD for the mean functions $f_1 - f_3$ for the generalized likelihood ratio procedure.
Figure 7.4: (b) The histograms of IMSE and PMAD for the mean functions $f_4 - f_6$ for the generalized likelihood ratio procedure.
Figure 7.5: (a) The histograms of IMSE and PMAD for the mean functions $f_1 - f_3$ for the fixed $\delta$ Bayesian procedure.
Figure 7.5: (b) The histograms of IMSE and PMAD for the mean functions $f_4 - f_6$ for the fixed $\delta$ Bayesian procedure.
Figure 7.6: (a) The histograms of IMSE and PMAD for the mean functions $f_1 - f_3$ for the first modified Bayesian procedure.
Figure 7.6: (b) The histograms of IMSE and PMAD for the mean functions $f_4 - f_6$ for the first modified Bayesian procedure.
Figure 7.7: (a) The histograms of IMSE and PMAD for the mean functions $f_1 - f_3$ for the second modified Bayesian procedure.
Figure 7.7: (b) The histograms of IMSE and PMAD for the mean functions $f_4 - f_6$ for the second modified Bayesian procedure.
Table 7.10: Simulation results for the counter example.

From the histograms for the IMSE and the PMAD, we can see that the pseudo-sequential procedures have larger IMSE and PMAD. This can be expected because the estimate of the location in these procedures overshoot the actual value. Both the modifications overcome this problem and are comparable to the procedures based on the ‘entire’ sample. The second modification which uses the data between the estimate of the location of the change-point and the stopping time as a training sample performs very well and in most cases better than all other procedures.

We also ran simulations using the mean function given by the counter example for the binary segmentation procedure. We ran simulations for \( n = 50, 100, 150, 200 \)
CHAPTER 7. NUMERICAL RESULTS

and 250. For each $n$ generated 2500 samples for the corresponding mean function. We used the mixture version of both the pseudo-sequential procedures and the fixed $\delta$ version of the Bayesian method with $\delta = 1$. Table 7.10 gives the number of change-points detected. We can see very clearly that the binary segmentation procedure fails to detect any change-point very often and the probability increases with $n$.

7.2 Factorial Experiments

We can see from the simulation results that, in small samples, the procedures generally under-estimate the number of change-points. There are various factors that contribute to this. We will now examine the effects of the factors.

If the number of change-points increases then the probability of underestimation increases, because there are more change-points to miss. But if the distance between change-points increases then the procedure has more data to distinguish between the two parts of the data. So here the probability of underestimation decreases. Similarly, if the magnitude of the differences in the means before and after a change-point is large then it is easier to distinguish between the two sets and hence under estimation is less probable. Finally, if two consecutive change-points have the same signs for the difference then it is easier to distinguish the first mean from the third mean, even if we miss the first change-point, where as if the two changes tend to nullify each other then detecting the second becomes more difficult if we miss the first change-point. We have conducted a factorial experiment which takes into account all these observations. It is a $2^4$ experiment with the factors and levels given in Table 7.11.

We have included in our factorial experiment the binary segmentation procedure, the generalized likelihood ratio procedure and both the fixed $\delta$ and the mixture version of the Bayesian method. We have not included the procedure based on the Schwarz criterion, because it is computationally infeasible.
Table 7.11: The four factors and the two levels of each of the factors we use for the experiment.

For any given procedure, the observations are obtained as follows. For every combination of the factors ran 1000 repetitions of the estimation procedure. We take the mean of the deviations of the estimated number of change-points from the true number and standardize it by the sample standard deviation. Then we will have observations that are approximately normal with unit variance. The simulation results are given in Tables 7.12 through 7.16. For each combination of the levels, the tables give the number of times the corresponding procedure detected $i$ change-points for $i = 0, \ldots, 9$. A standard analysis of this factorial experiment gives us the main effect of each of the factors as well as the various interactions.
<table>
<thead>
<tr>
<th>Factors</th>
<th>Number of Change-Points Detected</th>
<th>Normalized Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3 4 5 6 7 8 9</td>
<td></td>
</tr>
<tr>
<td>N D J S</td>
<td>0 0 0 0</td>
<td>231 435 294 40 0 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>0 0 0 1</td>
<td>0 200 748 52 0 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>0 0 1 0</td>
<td>0 0 20 975 5 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>0 0 1 1</td>
<td>0 0 31 968 1 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>0 1 0 0</td>
<td>0 28 109 310 546 7 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>0 1 0 1</td>
<td>0 1 456 535 7 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>0 1 1 0</td>
<td>0 0 0 974 26 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>0 1 1 1</td>
<td>0 0 0 967 33 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>1 0 0 0</td>
<td>0 181 355 304 140 17 3 0 0 0</td>
</tr>
<tr>
<td></td>
<td>1 0 0 1</td>
<td>0 140 756 102 2 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>1 0 1 0</td>
<td>0 0 0 11 119 862 8 0 0 0</td>
</tr>
<tr>
<td></td>
<td>1 0 1 1</td>
<td>0 0 0 121 877 2 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>1 1 0 0</td>
<td>0 17 54 148 245 290 244 2 0 0 0</td>
</tr>
<tr>
<td></td>
<td>1 1 0 1</td>
<td>0 0 2 169 597 230 2 0 0 0</td>
</tr>
<tr>
<td></td>
<td>1 1 1 0</td>
<td>0 0 0 0 0 974 26 0 0 0</td>
</tr>
<tr>
<td></td>
<td>1 1 1 1</td>
<td>0 0 0 0 0 965 33 2 0 0</td>
</tr>
</tbody>
</table>

Table 7.12: Factorial Experiment - Simulation results for the Bayesian method with $\delta = 1$. 
## Chapter 7. Numerical Results

<table>
<thead>
<tr>
<th>Factors</th>
<th>Number of Change-Points Detected</th>
<th>Normalized Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>N D J S</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0</td>
<td>351 459 172 18 0 0 0 0 0 0</td>
<td>-89.3594</td>
</tr>
<tr>
<td>0 0 0 1</td>
<td>0 356 621 23 0 0 0 0 0 0</td>
<td>-81.3683</td>
</tr>
<tr>
<td>0 0 1 0</td>
<td>0 4 61 930 5 0 0 0 0 0</td>
<td>-7.2474</td>
</tr>
<tr>
<td>0 0 1 1</td>
<td>0 0 73 927 0 0 0 0 0 0</td>
<td>-8.8696</td>
</tr>
<tr>
<td>0 1 0 0</td>
<td>30 160 399 407 4 0 0 0 0 0</td>
<td>-31.2015</td>
</tr>
<tr>
<td>0 1 0 1</td>
<td>0 13 610 375 2 0 0 0 0 0</td>
<td>-39.1457</td>
</tr>
<tr>
<td>0 1 1 0</td>
<td>0 0 0 986 14 0 0 0 0 0</td>
<td>3.7662</td>
</tr>
<tr>
<td>0 1 1 1</td>
<td>0 0 0 983 17 0 0 0 0 0</td>
<td>4.1565</td>
</tr>
<tr>
<td>1 0 0 0</td>
<td>310 400 223 62 4 1 0 0 0 0</td>
<td>-136.9180</td>
</tr>
<tr>
<td>1 0 0 1</td>
<td>0 0 303 652 44 1 0 0 0 0</td>
<td>-133.6377</td>
</tr>
<tr>
<td>1 0 1 0</td>
<td>0 0 5 35 167 787 6 0 0 0</td>
<td>-14.2556</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>0 0 0 1 212 787 0 0 0 0</td>
<td>-16.3950</td>
</tr>
<tr>
<td>1 1 0 0</td>
<td>23 73 211 308 275 110 0 0 0 0</td>
<td>-51.2134</td>
</tr>
<tr>
<td>1 1 0 1</td>
<td>0 0 3 280 613 104 0 0 0 0</td>
<td>-62.0184</td>
</tr>
<tr>
<td>1 1 1 0</td>
<td>0 0 0 0 989 11 0 0 0 0</td>
<td>3.3333</td>
</tr>
<tr>
<td>1 1 1 1</td>
<td>0 0 0 0 986 14 0 0 0 0</td>
<td>3.7662</td>
</tr>
</tbody>
</table>

Table 7.13: Factorial Experiment - Simulation results for the mixture version of the Bayesian method.
<table>
<thead>
<tr>
<th>Factors</th>
<th>Number of Change-Points Detected</th>
<th>Normalized Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>N D J S</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0</td>
<td>344 400 209 47 0 0 0 0 0 0</td>
<td>-75.0256</td>
</tr>
<tr>
<td>0 0 0 1</td>
<td>0 340 596 63 1 0 0 0 0 0</td>
<td>-70.0056</td>
</tr>
<tr>
<td>0 0 1 0</td>
<td>0 1 30 946 23 0 0 0 0 0</td>
<td>-1.1923</td>
</tr>
<tr>
<td>0 0 1 1</td>
<td>0 0 30 941 29 0 0 0 0 0</td>
<td>-0.1301</td>
</tr>
<tr>
<td>0 1 0 0</td>
<td>69 196 347 374 13 1 0 0 0 0</td>
<td>-30.9793</td>
</tr>
<tr>
<td>0 1 0 1</td>
<td>0 30 589 372 9 0 0 0 0 0</td>
<td>-36.4255</td>
</tr>
<tr>
<td>0 1 1 0</td>
<td>0 0 0 920 77 3 0 0 0 0</td>
<td>9.1550</td>
</tr>
<tr>
<td>0 1 1 1</td>
<td>0 0 0 934 64 2 0 0 0 0</td>
<td>8.2801</td>
</tr>
<tr>
<td>1 0 0 0</td>
<td>304 338 236 101 16 5 0 0 0 0</td>
<td>-113.0675</td>
</tr>
<tr>
<td>1 0 0 1</td>
<td>0 0 256 646 95 3 0 0 0 0</td>
<td>-116.9893</td>
</tr>
<tr>
<td>1 0 1 0</td>
<td>0 0 2 11 88 854 42 3 0 0</td>
<td>-4.8134</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>0 0 0 0 0 0 88 870 42 0 0 0</td>
<td>-4.0656</td>
</tr>
<tr>
<td>1 1 0 0</td>
<td>51 133 217 262 223 112 2 0 0 0</td>
<td>-50.5441</td>
</tr>
<tr>
<td>1 1 0 1</td>
<td>0 0 5 331 545 118 1 0 0 0</td>
<td>-59.2562</td>
</tr>
<tr>
<td>1 1 1 0</td>
<td>0 0 0 0 0 0 919 78 3 0 0</td>
<td>9.2186</td>
</tr>
<tr>
<td>1 1 1 1</td>
<td>0 0 0 0 0 0 901 97 2 0 0</td>
<td>10.3681</td>
</tr>
</tbody>
</table>

Table 7.14: Factorial Experiment - Simulation results for the generalized likelihood ratio method.
## Table 7.15: Factorial Experiment - Simulation results for the binary segmentation procedure.
Table 7.16: Factorial Experiment - The main effects and the interactions for the four procedures.
Figure 7.8: Q-Q plot of the effects for Bayesian method - $\delta = 1$. 
Figure 7.9: Q-Q plot of the effects for Bayesian method - Mixture.
Figure 7.10: Q-Q plot of the effects for Likelihood Ratio method.
Figure 7.11: Q-Q plot of the effects for Binary Segmentation Procedure.
CHAPTER 7. NUMERICAL RESULTS

From the factorial experiment we conclude the following. The shape of the mean function has a significant effect on the binary segmentation procedure. A monotonic function makes it easier for the binary segmentation procedure to detect change-points and it also overestimates the number of change-points. The effect of the shape of the mean function is not significant in any of the pseudo-sequential procedures. The most important factor in all the procedures is the size of the jump in the mean. The more the difference the easier it is for the procedure to detect the change-point. The distance between the change-points is the next important main effect. All the procedures find it easier to detect the change-points if the distances between the change-points increase. The main effect corresponding to the number of change-points is also significant: a larger number makes it difficult to detect the change-points. Two of the two factor interactions NJ and DJ are significant and ND is marginally significant in all the procedures. The interaction JS is significant in the binary segmentation procedure. All other higher order interactions are not significant.

7.3 The Nile Data

We will now use the different procedures in Table 7.1 to estimate the number of change-points and their locations in the Nile data (Cobb (1978)). The data gives the annual volume (in $10^8$ m$^3$) of the Nile river from the years 1871 to 1970. We would like to detect any change in the average volume over the 100 years. We use the estimate of standard deviation (125) given in Cobb to transform the data into a unit variance data. The graphs in Figures 7.9 (a)-(d) give the likelihood functions for the various procedures we described. we used a five percent critical level for all the test procedures. The likelihood function for the pseudo-sequential procedures starts afresh every time the procedure crosses the boundary. For the procedure based on the Schwarz criterion, Table 7.17 gives the minimal sum of squares and the Schwarz
<table>
<thead>
<tr>
<th>$r$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS($r$)</td>
<td>2835112.00</td>
<td>1597408.00</td>
<td>1542280.00</td>
<td>1438077.25</td>
<td>1341810.00</td>
</tr>
<tr>
<td>SC($r$)</td>
<td>512.621</td>
<td>488.541</td>
<td>491.390</td>
<td>492.498</td>
<td>493.639</td>
</tr>
<tr>
<td>SS($r$)</td>
<td>181.449</td>
<td>102.2363</td>
<td>98.7079</td>
<td>92.0390</td>
<td>85.8782</td>
</tr>
<tr>
<td>SC($r$)</td>
<td>29.7901</td>
<td>5.7110</td>
<td>8.5601</td>
<td>9.6677</td>
<td>10.8087</td>
</tr>
</tbody>
</table>

Table 7.17: The Problem of Nile - The minimal sum of squares and Schwarz criterion for both the original data and the transformed data.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Number</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS</td>
<td>1</td>
<td>1898</td>
</tr>
<tr>
<td>SC</td>
<td>1</td>
<td>1898</td>
</tr>
<tr>
<td>BM</td>
<td>1</td>
<td>1902</td>
</tr>
<tr>
<td>LR</td>
<td>1</td>
<td>1902</td>
</tr>
<tr>
<td>BM - $\delta$</td>
<td>1</td>
<td>1903</td>
</tr>
<tr>
<td>BM - M1</td>
<td>1</td>
<td>1898</td>
</tr>
<tr>
<td>BM - M2</td>
<td>1</td>
<td>1898</td>
</tr>
</tbody>
</table>

Table 7.18: The Problem of Nile - The estimated number and the locations of the change-points for all the procedures used.

criterion for $r = 0, 1, 2, 3, 4$, both for the original data and the transformed data. Since the Schwarz criterion is minimized for $r = 1$, by Remark 3.1, we get that the minimal sum of squares is attained at year 1898. The number of change-points detected and the estimates of their locations are given in Table 7.18.

We can see that all the procedures detect exactly what has been known before. The value of the stopping times for the pseudo-sequential procedures are not good estimates of the location of the change-point. But we can estimate the location using the maximum likelihood within the segment where the change is detected and do as well as the other procedures.
Figure 7.12: (a) The Problem of Nile - Data and the likelihood function for the binary segmentation procedure.
Figure 7.12: (b) The Problem of Nile - The likelihood functions for the Bayesian mixture method and the generalized likelihood ratio method.
Figure 7.12: (c) The Problem of Nile - The likelihood function for the fixed $\delta$ Bayesian method and the second modified mixture Bayesian method.
Figure 7.12: (d) The Problem of Nile - The likelihood function for the two segments from the binary segmentation procedure and the data until the value of the stopping time for the modified mixture Bayesian methods.
Chapter 8

Exponential Processes

In this chapter we will consider a similar problem, that is, the detection and estimation of change-points in exponential processes. Here again we will be dealing with a multiple change-point problem. We want to estimate the number and the locations of the changes in the rate of an exponential process. The problem can be stated as follows. Let \( \{X_i^{(n)}; 1 \leq i \leq n < \infty \} \) be independent exponential random variables with rate \( \lambda_i^{(n)} \). Let \( 0 = \nu_0^{(n)} < \nu_1^{(n)} < \ldots < \nu_m^{(n)} < \nu_{m+1}^{(n)} = n \) be the change-points, that is, there exists a sequence of constants \( \{\theta_j; j \geq 0\} \) such that \( \theta_j \neq \theta_{j+1} \) for all \( j \) and

\[
\lambda_i^{(n)} = \theta_j, \text{ if } \nu_j^{(n)} < i \leq \nu_{j+1}^{(n)}, j = 0, \ldots, m_n. \tag{8.1}
\]

We are interested in estimating the number (\( m_n \)) of change-points and their locations (\( \nu_j^{(n)}, j = 1, \ldots, m_n \)) as well as the rates \( \theta_j, j = 0, \ldots, m_n \). See Loader (1990) for a more extensive treatment of point processes.

8.1 Consistency Result

We would like derive the Shiryaev-Roberts type procedures for this case. Pollack and Siegmund (1991) used invariance arguments to detect changes in the mean of a
Gaussian process, when the initial mean is unknown. We would like to use similar
invariance arguments to detect changes in the rate function when the initial rate
is unknown. Observe that the problem is scale invariant. We will now derive the
likelihood statistic based on the scale invariant function for the Bayesian (pseudo-
sequential) method.

Let $U_1, \ldots, U_n$ be independent exponential random variables. Then the scale
invariant function is $\{V_i; 2 \leq i \leq n\}$ where $V_i = U_i/U_1$. We want to derive the
likelihood ratio statistic based on the $V$'s to test the hypothesis that all the $U$'s have
the same rates against the alternative that there is a $k$ and constant $\delta > 0$ such
that $U_1, \ldots, U_k$ have rate $\theta$ and $U_{k+1}, \ldots, U_n$ have rate $\delta \theta$. Since the function is scale
invariant we can assume without loss of generality that $\theta = 1$. The likelihood of
$V_2, \ldots, V_n$ under the null is

$$
\int_0^\infty v_1^{n-1} \exp \left\{ -v_1 \left( 1 + \sum_{i=2}^n v_i \right) \right\} \, dv_1 = \left( 1 + \sum_{i=2}^n v_i \right)^n
$$

and under the alternative is

$$
\int_0^\infty \delta^{-k} v_1^{n-1} \exp \left\{ -v_1 \left( 1 + \sum_{i=2}^k v_i + \delta \sum_{i=k+1}^n v_i \right) \right\} \, dv_1
$$

$$
= \delta^{-k} \left( 1 + \sum_{i=2}^k v_i + \delta \sum_{i=k+1}^n v_i \right)^n,
$$

where $V_1 = U_1$ and $v_i$'s are the values of the random variables $V_i$'s. So the likelihood
ratio test statistic is

$$
\delta^{-k} \left( 1 + \sum_{i=2}^n V_i \right)^n / \left( 1 + \sum_{i=2}^k V_i + \delta \sum_{i=k+1}^n V_i \right)^n,
$$

which is the same as

$$
\delta^{-k} \left( \sum_{i=1}^n U_i \right)^n / \left( \sum_{i=1}^k U_i + \delta \sum_{i=k+1}^n U_i \right)^n.
$$
CHAPTER 8. EXPONENTIAL PROCESSES

From this derivation of the likelihood ratio based on the invariant function we define the analogue of the statistic in (5.1) as

\[ R_n(\delta) = \sum_{k=0}^{n-1} \delta^{n-k}[S_n/(S_k + \delta(S_n - S_k))]^n, \]

where \( S_i = X_1 + \cdots + X_i \). So we will use the statistic

\[ S(X_1, \ldots, X_k) = (R_k(\delta) + R_k(1/\delta))/2 \]

(8.2)

to define the stopping rules for the pseudo-sequential method. The estimates of the number of change-points \( \hat{m}_n \) and their locations \( \hat{\nu}_j^{(n)} \) are as given in (4.2). In order to prove the consistency result we need the following lemma.

Lemma 8.1 Let \( f(\delta, \epsilon) = [\delta + (1-\delta)\epsilon]/\delta^{1-\epsilon} \) on the unit square. Then

\[ f(\delta, \epsilon) \geq \exp\{(1-\delta)^2\epsilon(1-\epsilon)/2\}. \]

Proof: We will prove the result by showing that the logarithm of both functions satisfy the inequality.

\[
\log f(\delta, \epsilon) = \log(\delta + (1-\delta)\epsilon) - (1-\epsilon)\log \delta \\
\leq -\sum_{k=1}^{\infty} \left[ (1-\delta)(1-\epsilon) \right]^{k} / k + (1-\epsilon) \sum_{k=1}^{\infty} \frac{(1-\delta)^k}{k} \\
= \sum_{k=1}^{\infty} \frac{(1-\delta)^k}{k} [(1-\epsilon) - (1-\epsilon)^{k}] \\
\geq \frac{(1-\delta)^2}{2} [(1-\epsilon) - (1-\epsilon)^2],
\]

where the second inequality is from the expansion \( \log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots \), applied to \( \log(\delta + (1-\delta)\epsilon) \) and \( \log \delta(\delta + (1-\delta)\epsilon = 1 - (1-\delta)(1-\epsilon) \) and \( \delta = 1 - (1-\delta) \). The rest of the proof is simple algebra. \( \Box \)

We will now prove the following consistency statement. The statement and the proof follow the prototype in Section 4.2.
CHAPTER 8. EXPONENTIAL PROCESSES

Theorem 8.1 Let $X_{i}^{(n)}$ be as defined in (8.1). Assume that

(i) $\inf_{i} \{ \nu_{i+1}^{(n)} - \nu_{i}^{(n)} \} \geq c_n \log n \ \forall n$, where $c_n \to \infty$ as $n \to \infty$

(ii) $\inf_{i} | \log \theta_{i+1} \log \theta_{i} | \geq - \log \delta$, for some $\delta < 1$ independent of $n$.

Let $N_{i}^{(n)}$ be the stopping rules defined in (4.1) using the statistic defined in (8.2). Let $\beta_n = d_n \log n$ for some $d_n$ such that $d_n \to \infty$ and $d_n/c_n \to 0$ as $n \to \infty$. Let $b_n = n^2$ and let

$$\mathcal{A}_n = \left\{ \nu_{i}^{(n)} < N_{i}^{(n)} < \nu_{i}^{(n)} + \beta_n, 1 \leq i \leq m_n \& N_{m_n+1}^{(n)} = \infty \right\}.$$ 

Then $P(\mathcal{A}_n)$ converges to 1 as $n$ goes to infinity.

Proof: As in the Gaussian case, we will use the following setup. Let $U_1, U_2, \ldots$ be independent exponential random variables. Let $P_{\infty}$ denote the measure under which all the $U_i$'s have unit rate. Let $P_{\nu}^{(\lambda)}$ denote the measure under which $U_1, \ldots, U_{\nu}$ have rate 1 and $U_{\nu+1}, U_{\nu+2}, \ldots$ have rate $\lambda$. Let $V_k = U_1 + \cdots + U_k$ and let

$$N = \inf \{ k : S(U_1, \ldots, U_k) \geq b \},$$

where $S$ is as in (8.4) and $b$ as in the theorem. Let $B_{in}$ and $C_{in}$ be as defined in (4.7) and (4.8) respectively. By appealing to the general consistency theorem and the method of proof given in Section 4.2, it is enough to get good upper bounds for $P_{\infty}(N \leq n)$ and $\max_{\beta_n \leq \nu \leq m_n - \beta_n} P_{\nu}^{(\lambda)}(N > \nu + \beta_n)$.

Since the statistic is the likelihood ratio, by the martingale properties, we get that $P(N \leq n) \leq n/b_n = 1/n$. Thus

$$\sum_{i=0}^{m_n} P(B_{in} \leq (m_n + 1)P(N \leq n) \leq 1/\log n. \quad (8.3)$$

This bounds the probability of detecting a non-existent change-point.

The calculation of the probability of missing a change-point is more complex. We want to find $P_{\nu}^{(\lambda)}(N \geq \nu + 1)$ for $\nu \geq 1$. We can assume without loss of generality
that $\lambda < 1$. Then

$$
P^{(\lambda)}_{\nu}(N > \nu + \beta_n) \leq P^{(\lambda)}_{\nu}(S(U_1, \ldots, U_k) < n^2) \leq P^{(\lambda)}_{\nu}(R_{\nu+\beta_n}(\delta) < 2n^2) \leq P^{(\lambda)}_{\nu}\left(\delta^l \left[\frac{V_{\nu+\beta_n}}{V_{\nu} + \delta(V_{\nu+\beta_n} - V_{\nu})}\right]^k < 2n^2\right) \leq n^{-d_n^{1/3}/2}.
$$

The first three inequalities are from the definition of $N$, $S(U_1, \ldots, U_k)$ and $R_k(\delta)$ respectively and the final inequality if from Lemma 8.2. Thus

$$
\sum_{i=1}^{m_n} P(C_{i,n}) \leq \sum_{i=1}^{m_n} \max_{\beta_n \leq \nu \leq \beta_m} P^{(\theta_i/\theta_i-1)}_{\nu}(N > \nu + \beta_n) \leq \frac{n}{\log n} n^{-d_n^{1/3}/2}.
$$

(8.4)

By equations (8.3) and (8.4), the bounds for the probabilities of errors of both kind of errors go to zero. Hence the theorem follows. \hfill \textbf{Q.E.D.}

\textbf{Lemma 8.2} Let $l = d_n \log n$, $l \leq \nu \leq n$ and $k = \nu + l$. Let $U$ and $V$ be independent with $\text{Gamma}(\nu, 1)$ and $\text{Gamma}(l, 1)$ distribution respectively. Let $0 < \lambda \leq \delta < 1$. Then

$$
P\left\{\delta^l \left(\frac{U + V}{\lambda}\right)^k < 2n^2\right\} \leq n^{-d_n^{1/3}/2}.
$$

\textbf{Proof:} Let $W = U/V$ and $Z_\lambda = (U + V)/\lambda/(U + \delta V)/\lambda$. Then $Z_\lambda = (W + 1/\lambda)/(W + \delta/\lambda) = (\lambda W + 1)/(\lambda W + \delta)$. Observe that, for $0 < \delta < 1$, $g(x) = (x + 1)/(x + \delta)$ is a decreasing function in $x$, on $[0, \infty)$. So $Z_\lambda$ is stochastically decreasing. So it is enough to prove that $P\{\delta^l Z_\delta^k < 2n^2\} \leq n^{-d_n^{1/3}}$.

Let $l_1 = e_n \log n$, where $e_n = d_n^{2/3}$, and $\eta = (l - l_1)/k$. Let $B = V/(U + V)$. Observe that on the set \{\text{Set} \{B \geq \eta\}\},

$$
\delta^l Z_\delta^k = \delta^l \{(1 + (1/\delta - 1)B)^k \geq \delta^l \{1 + \eta(1 - \delta)/\delta\}^k.
$$

(8.5)
By Lemma 8.1
\[
\delta^l \left(1 + \eta \left(1 - \frac{1 - \delta}{\delta}\right)^k\right) = \delta^{l_1} \left(\delta^\eta \left(1 + \eta \left(1 - \frac{1 - \delta}{\delta}\right)^k\right)\right)
\]
\[
= \delta^{l_1} \left(\delta + \eta \left(1 - \frac{1 - \delta}{\delta^{1-\eta}}\right)^k\right)
\]
\[
\geq \delta^{l_1} \exp\left(k(1 - \delta)^2 \eta (1 - \eta) / 2\right).
\]

Observe that, \(1 - \eta \geq 1/2\). Also \(\eta k = l - l_1 = (d_n - d_n^{2/3}) \log n\), which for large \(n\) is larger than \(d_n \log n / 2\), since \(d_n \to \infty\). So, for large \(n\)
\[
\delta^{l_1} \exp\left(k(1 - \delta)^2 \eta (1 - \eta) / 2\right) \geq \delta^{e_n \log n} \exp\left((1 - \delta)^2 d_n \log n / 8\right)
\]
\[
= \exp\left\{(e_n \log \delta \log n) + ((1 - \delta)^2 d_n \log n / 8)\right\}
\]
\[
= n^{(e_n \log \delta) + ((1 - \delta)^2 d_n / 8)}.
\] (8.6)

Since \(e_n\) goes to infinity at a slower rate than \(d_n\), the exponent in (8.6) goes to infinity. So by (8.5) and (8.6) we get that \(\delta^l Z_\delta^k\) is larger than \(2n^2\) on the set \(\{B \geq \eta\}\). Thus \(\{\delta^l Z_\delta^k \leq 2n^2\}\) is contained in the set \(\{B \leq \eta\}\).

Now we have to find \(P(B \leq \eta)\). We will use the relation between the binomial and the beta function. Since \(B\) has a beta distribution with \(l\) and \(\nu\) degrees of freedom
\(P(B \leq \eta) = 1 - B(l - 1, k - 1, \eta)\), where \(B(i, m, p)\) is the distribution function of a Binomial\((m, p)\) random variable (e.g. Feller (1968, Vol. 1, p.173)). Let \(E_1, E_2, \ldots\) be independent bernoulli trials with probability of success \(\eta\). Let \(F_i = E_1 + \cdots + E_i\). We want \(P(F_{k-1} \geq l)\). We will use the moment generating function of the \(E_i\)’s to bound this probability. Let \(\psi(\lambda) = \log E(\exp \lambda E_1)\). So \(\psi(\lambda) = \log((1 - \eta) + \eta \exp \lambda)\).

Let \(t = l_1 / (k - 1)\). Then observe that \(\eta \leq t\) for large \(n\). So for large \(n\), and all \(\lambda\),
\[
P(F_{k-1} \geq l_1) \leq \exp\{(k - 1)\psi(\lambda) - \lambda t\}\). We want to minimize \(\psi(\lambda) - \lambda t\) over \(\lambda\). We can see that the minimum is attained \(\lambda = t(1 - \eta) / (1 - t)\).
So

\[
P\{F_{k-1} \geq (k-1)t\} \\
\leq \exp\left\{ (k-1) \left[ \log \left( \eta \left[ \frac{t(1-\eta)}{\eta(1-t)} \right] + (1-\eta) \right) - t \log \left[ \frac{t(1-\eta)}{\eta(1-t)} \right] \right] \right\} \\
= \exp\left\{ (k-1) \left[ (1-t) \log \left( \frac{1-\eta}{1-t} \right) + t \log \left( \frac{\eta}{t} \right) \right] \right\}.
\]

Now use \(\log(1+x) \leq x\) and \(\log(1-x) \leq -x - x^2/2\) for \(x > 0\) to get

\[
\exp\left\{ (k-1) \left[ (1-t) \log \left( \frac{1-\eta}{1-t} \right) + t \log \left( \frac{\eta}{t} \right) \right] \right\} \\
\leq \exp\left\{ (k-1) \left[ (1-t) \left( \frac{t-\eta}{1-t} \right) + t \left( -\frac{t-\eta}{t} - \frac{(t-\eta)^2}{2t^2} \right) \right] \right\} \\
= \exp\{-l_2^2/(2t)\}
\]

The final inequality follows from \((k-1)(t-\eta) = (k-1)t - k\eta + \eta = l - (l_1 - \eta) + \eta \geq l_1\).

Since \(l_2^2/l = c_n^2 \log n/d_n = d_n^{1/3} \log n\) we get that

\[
\exp\{-l_2^2/(2l)\} \leq n^{-d_n^{1/3}/2}.
\]

This completes the proof of the lemma and hence the theorem. \(\square\)

We can get the following corollaries which deals with the consistent estimation of the number and locations as well as the rates from this theorem. The proofs are similar to the ones in Gaussian case.

**Corollary 8.1** Under the conditions of Theorem 8.1 the following results are true.

(i) \(P(\hat{m}_n = m_n) \rightarrow 1\) as \(n \rightarrow \infty\),

(ii) \(|\hat{\nu}_i^{(n)}(\hat{m}_n) - \nu_i^{(n)}|/n\) converges to 0 with probability one.

**Corollary 8.2** Let \(\hat{\theta}_j = (S_{j+1} - S_j)^{-1}(\hat{\nu}_{j+1} - \hat{\nu}_j), 0 \leq j \leq \hat{m}_n, \) where \(\hat{\nu}_0 = 0\) and \(\hat{\nu}_{\hat{m}_n+1} = n\). Then, under the conditions of Theorem 8.1, \(\hat{\theta}_j\) converges to \(\theta_j\) in probability.
CHAPTER 8. EXPONENTIAL PROCESSES

The theorem we proved deals with the fixed $\delta$ case, where $\delta$ is the largest value the ratio of the smaller to the larger of the rates of the exponential random variables before and after the change-point. The procedure requires the knowledge of $\delta$. We can define a mixture procedure by using a prior for $\delta$ and prove a similar consistency result. We can also use the likelihood ratio statistics, which would be the standard $F$-statistic, to define the stopping rules and get a similar consistency result.

This theorem can be easily generalized to the case of general Gamma variables where the scale parameter changes. A consistency result for Gamma variables with shape parameter $1/2$ would give a procedure for estimating changes in the variance of a Gaussian process when the mean is known.

8.2 British Coal-Mining Disaster Data

We will now use the method we developed, on the British coal-mining disaster data. This data covers 191 mining explosions from 15 March, 1851 to 22 March, 1962. The time in days between consecutive explosions are reported (see for example Jarrett (1979)). We would like to see if there is a change in the rate of the explosions. We will use the procedure we described with $\delta = 0.5$ on the data to detect any change-point that might be present. The data and the Bayesian likelihood function given by (8.2) for $\delta = 0.5$ are given in Figure 8.1. We use a 5 percent critical value to test for a change. The stopping rule with that value of $b$ stops after the 134$^{th}$ observation and we start looking for the next change from observation number 135 onwards. The likelihood function reflects it by a spike at 134. The procedure detects a single change-point only and it is before the 134$^{th}$ observation. We can estimate the exact location using the mode of the posterior probability distribution for the location or some similar method. See Loader (1990) and references there for a discussion on this data set.
CHAPTER 8. EXPONENTIAL PROCESSES

British Coal-Mining Disaster Data

Bayesian Likelihood Function

Figure 8.1: The British coal-mining disaster data and the Bayesian likelihood function to detect a change in the rate.
Chapter 9

Summary

In the introduction we described the general multiple change-point problem. In Chapters 2 through 7 we restricted ourselves to the Gaussian process with constant variance where we are interested in detecting changes in the mean and estimating their locations.

We dealt with two classes of procedures that could be used for this problem. One is the class of procedures based on the 'entire' sample and the other is the class of pseudo-sequential procedures. We considered two members from the former class which are the binary segmentation procedure and the procedure based on the Schwarz criterion. Both these procedures have been discussed before and we proved the consistency of both the procedures under weaker conditions on the number and the locations of the change-points than before. We also showed, by an example, that the necessary conditions for the binary segmentation procedure to be consistent is much stronger than the sufficient conditions for the other procedures.

The estimates of both these depend on the entire rather than the part of the sample that is close to a given change-point. So we introduced the pseudo-sequential procedures which use the sample sequentially to detect the change-points and estimate their locations. We described the procedure for the general problem and gave a
prototype of the consistency statement and the method of proof for this class of procedures. For the Gaussian case, we considered two kinds of pseudo-sequential procedures. One is the Bayesian method motivated by the Shirayev-Roberts scheme and the other is the likelihood ratio method motivated by the CUSUM tests. We proved the consistency of both these procedures.

The procedure based on the Schwarz criterion is computationally infeasible even for moderate data size and five or more change-points. The estimates of the pseudo-sequential procedures are not affected by the observations in the future and not much by the observations in the distant past. This enables us to estimate the number and the locations of the change-points more efficiently. The problem of overestimating the location of the detected change-point can be overcome by estimating the location of the change in the segment between consecutive stopping times either by the posterior probability distribution of the location or by maximum likelihood estimate of the location.

The Monte Carlo results show us that the binary segmentation procedure is affected by the shape of the mean function. It also has a severe overestimation as well as underestimation problem. As mentioned earlier the procedure based on Schwarz criterion is computationally infeasible and it too suffers from an overestimation problem. The pseudo-sequential procedures as described in Chapter 4, underestimate the number of change-points if there isn’t enough data to distinguish between the different means. It also overestimates the locations of the change-points. The modified procedures where we estimate the location of the change-point, instead of using the value of the stopping time itself performs much better. The use of training sample further increases the efficiency of the procedure. The Bayesian method, since it is not affected much by outliers, performs better than the likelihood ratio method. While the likelihood ratio method can be easily generalized, the Bayesian method cannot be. So we would prefer to use the pseudo-sequential procedures.
CHAPTER 9. SUMMARY

We also discussed the Bayesian method for exponential processes and proved the consistency of the procedure. We used it on the British coal-mining disaster data to detect a change in the rate of the accidents. The procedure can be generalized to gamma random variables and can be used to detect changes in the variance of a Gaussian process.

There are other methods that have been studied. One common procedure is to use a moving window along the data to detect change-points with either likelihood ratio statistics or nonparametric statistics. These are similar to the pseudo-sequential procedures in concept, but are modified to detect changes with fewer computations. A drawback with a fixed window width procedure is that it has to wait too long to detect a drastic change in the distribution. We can fix it by using an upper bound for the window size and try all window sizes up to that bound. If we set the bound to be the whole data size the we get back to the pseudo-sequential procedures. The choice of window size is crucial in these procedures. There doesn’t seem to be any natural way of choosing the window width.
Appendix A

Tail Probability Approximations

In order to run any of the procedure we need to control the error probabilities to be within a desirable limit. Since the probability of not detecting a change when there is one depends on the difference in mean, which is unknown, we can't do much about it. We can control the probability of detecting a change when there isn't any. This can be done by suitably choosing the cost function for the procedure based on the Schwarz criterion. For all the other procedures, since they use a test to detect a change-point, we can control this error probability, by choosing the critical value appropriately. For this we need the tail probability of the test statistic we use for the different procedures. We will now talk about the tail probability approximations we used to determine the critical value and derive some approximations.

A.1 Critical Values

Since we are dealing with the probability under no change we have the following setup. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. standard normal random variables and let $S_i = Y_1 + \cdots + Y_i$. Let $T(Y_1, \ldots, Y_n)$ be the test statistic we use. Then we need the tail probabilities for this statistic, i.e., $P\{T(Y_1, \ldots, Y_n) > b\}$, so that we can choose $b$ suitably. We will
now describe the approximations we used $P\{T(Y_1, \ldots, Y_n) > b\}$.

First, we will consider the binary segmentation procedure. Let $Z_i = (iS_i/n - S_i)/[(i(1 - i/n))^{1/2}$. We want $b$ such that $P\{\max_{1 \leq i \leq n-1} |Z_i| > b\}$ is small. This is a two-sided probability. Since $P\{\max_{1 \leq i \leq n-1} Z_i > b\}$ and $P\{\min_{1 \leq i \leq n-1} Z_i < -b\}$ is negligible, we can approximate the two sided probability, by adding the two one-sided probabilities $P\{\max_i Z_i > b\}$ and $P\{\min_i Z_i < -b\}$. Siegmund (1985) gives the following approximation for this one-sided probability.

$$P\left\{ \max_{n_0 \leq i \leq n_1} Z_i \geq b \right\} \approx 1 - \Phi(b) + \phi(b) \int x^{-2}\nu(x + b^2/(nx))dx, \quad (A.1)$$

where

$$\nu(x) = 2x^{-2}\exp\left\{-2 \sum_{m=1}^{\infty} m^{-1}\Phi(-xm^{1/2}/2)\right\}, \quad (A.2)$$

$\phi$ and $\Phi$ are the standard normal density and cumulative distribution function respectively, and the integral is taken over the range $(b(n_1^{-1} - n^{-1})^{1/2}, b(n_0^{-1} - n^{-1})^{1/2})$.

Now, by setting $n_0 = 1$ and $n_1 = n - 1$, we get the tail probability approximation we need, which can then be inverted (numerically) to obtain the critical value. We can also derive this approximation using Woodroofe's method.

We will now consider the pseudo-sequential procedures. The pseudo-sequential procedures involve stopping rules of the type

$$N = \inf\{k : S(Y_1, \ldots, Y_k) > b\},$$

where $S$ is a suitably chosen statistic. We want to choose $b$ in such a way that $P\{N \leq n\} \leq \alpha$, for a given significance level $\alpha$. We will use the idea that the tail of stopping rule $N$ behaves approximately like exponential random variables. So we can find $P\{N \leq n\}$ using the equation

$$P\{N \leq n\} \approx 1 - \exp\{-n/E(N)\}. \quad (A.3)$$

So we are done if we find the expectation of $N$ for any given $b$. Pollack and Siegmund (1991) give an asymptotic expression for $E(N)$ if the stopping rule is defined using
the statistic $R_k(\delta)$ given by (5.1) and any mixture of it. The expected run lengths are $b/\nu(\delta)$ and $b/\int \nu(\delta)G(d\delta)$ respectively, where $\nu$ is the function defined by (A.2), $\delta$ is the parameter chosen and $G$ is the prior on $\delta$ for the mixture procedure. We can get the critical levels by inverting equation (A.3).

For the likelihood ratio method, we don't have asymptotic expressions for the expected run lengths. So we cannot use the idea behind equation (A.3). Let $Z_{ij}$ be the likelihood ratio statistic we use to test whether $i$ is a change-point in $Y_1, \ldots, Y_j$. Then by the definition of the statistic in the likelihood ratio case observe that, the set $\{N \leq n\}$ is same as $\{\max_{1 \leq i < j \leq n} Z_{ij} > b\}$. So we want an approximation for $P\{\max_{1 \leq i < j \leq n} Z_{ij} > b\}$. This is a maxima of a two dimensional random field. We can find this probability by using the method in Siegmund (1988). For the generalized likelihood ratio statistic, i.e. $Z_{ij} = |iS_j/j - S_i/[i(1-i/j)]|^{1/2}$, this method yields the approximation

$$P\left\{ \max_{1 \leq i < j \leq n} Z_{ij} > b \right\} \approx \frac{b^3 \phi(b)}{2} \int \int \frac{1}{(t-s)^2} \nu \left( b \sqrt{\frac{t}{s(t-s)}} \right) \nu \left( b \sqrt{\frac{s}{t(t-s)}} \right) dsdt,$$

(A.4)

where the function $\nu$ is defined by (A.2), $\phi$ is the standard normal density and the integral is taken over the set $\{(s,t) : s, n-t, t-s \geq 1\}$. We use this type of approximations to derive the expected run lengths of stopping rules using generalized likelihood ratio statistics for sequential detection of change-points (Siegmund and Venkatraman (to be published)). We will now derive the approximations for the binary segmentation procedure and the likelihood ratio procedure given by (A.1) and (A.4) respectively.

### A.2 Derivations

Let $Y_1, \ldots, Y_n$ be i.i.d standard normal random variables and let $S_i = Y_1 + \cdots + Y_i$. Let $Z_i = (iS_n/n - S_i)/[i(1-i/n)]^{1/2}$. Then approximations for $P\{\max_{1 \leq i \leq n-1} Z_i > b\}$,
given by (A.1), give us the critical values for the binary segmentation procedure. We will now sketch the derivation of this approximation. The idea of this approximation is to divide the event into disjoint subsets and approximate them. Then the sum, which can be replaced by an integral, gives us the approximation.

We will approximate the probability by dividing the event into subsets by looking at the last time \( Z_i > b \). Then, observe that \( P \{ \max_{1 \leq i \leq n-1} Z_i > b \} \) is given by

\[
P \{ Z_{n-1} > b \} + \sum_{i=1}^{n-2} P \left\{ \max_{i+1 \leq j \leq n-1} Z_j \leq b | Z_i > b \right\} = 1 - \Phi(b) \\
+ \sum_{i=1}^{n-2} \int_0^\infty \phi \left( b + \frac{x}{\sqrt{n}} \right) P \left\{ \max_{i+1 \leq j \leq n-1} \sqrt{n}(Z_j - Z_i) \leq -x | Z_i = b + \frac{x}{\sqrt{n}} \right\} \frac{dx}{\sqrt{n}}. \tag{A.5}
\]

In order to compute (A.5) we need the conditional probabilities within the integral. We will now calculate the conditional means and covariances of the random variables involved.

Observe that for \( j > i \)

\[
\text{Cov}(Z_i, Z_j) = \frac{i(n-j)}{j(n-i)}.
\]

So, \( E(Z_j | Z_i = z) = \sqrt{i(n-j)/j(n-i)}z \). Let \( \xi = b + x/\sqrt{n} \), \( s = i/n \) and \( b/\sqrt{n} \) converge to \( b_0 \). Then for all \( j - i \) of smaller order than \( n \)

\[
E \left\{ \sqrt{n}(Z_j - Z_i) | Z_i = \xi \right\} = \sqrt{n} \left[ \frac{i(n-j)}{j(n-i)} - 1 \right] \xi \\
= \sqrt{n} \left[ \frac{i(n-j) - j(n-i)}{\sqrt{j(n-i)} \left( \sqrt{j(n-i)} + \sqrt{i(n-j)} \right)} \right] \xi \\
\rightarrow -(j - i) \frac{1}{2s(1-s)} b_0. \tag{A.6}
\]

Similarly for \( i < j_1 < j_2 \), with \((j_2 - i)/n\) converging to zero,

\[
\text{Cov} \left\{ \sqrt{n}(Z_{j_1} - Z_i), \sqrt{n}(Z_{j_2} - Z_i) | Z_i = \xi \right\}
\]
From the calculations in (A.6) and (A.7) we get that the conditional probability in (A.5) converges to \( P\{\max_{j \geq 1} W_j < -x\} \) where \( W_j \) is a random walk with normal increments, with mean \(-b_0[2s(1-s)]^{-1}\) and variance \([s(1-s)]^{-1}\).

Let \( U_m \) be a normal random walk with \( N(-\mu, \sigma^2) \) increments, where \( \mu > 0 \). By Siegmund (1989),

\[
\int_0^\infty \exp(-2\mu x/\sigma^2) P\left\{ \max_{j \geq 1} U_j \leq -x \right\} \, dx = \mu \nu(2\mu/\sigma),
\]

where \( \nu \) is the function given by (A.2). Using (A.8) we get the integrals in the sum (A.5) reduce to

\[
\frac{b\phi(b)}{2s(1-s)\nu} \nu \left( \sqrt{\frac{1}{s(1-s)b_0}} \right),
\]

where \( s = i/n \) and \( b_0 = b/\sqrt{n} \). Now observe that the sum looks like a Riemann sum, and can be replaced by the following integral

\[
b\phi(b) \int \frac{1}{2s(1-s)} \nu \left( \sqrt{\frac{1}{s(1-s)b_0}} \right) \, ds,
\]

where the integral is over the range \((n^{-1}, 1-n^{-1})\). Finally a change of variable will give us the form in (A.1).

This approximation has been proved in Siegmund (1985, Chapter XI). The proof here is different. See James, James and Siegmund (1987) and the references there for numerical results.
We will now derive the approximation in (A.4). The approximation for the binary segmentation procedure deals with the maximum of an one dimensional random field. Here we will be dealing with the maxima of a two dimensional random field. Let the $Y_i$'s and the $S_i$'s be as before and let $Z_{ij} = (iS_j/j - S_i)/[i(1 - i/j)]^{1/2}$. We want $b$ such that $P(\max_{1 \leq i < j \leq n} |Z_{ij}| > b)$ is small. As before this two-sided probability can be approximated by the one-sided probability. We will now derive an approximation for the one-sided probability, $P(\max_{1 \leq i < j \leq n} Z_{ij} > b)$, using the method discussed in Siegmund (1988). We will break up the event that the process the boundary into disjoint subsets by looking at the last time the process crosses the level $b$. We will approximate the probability of each of the subsets and sum them (which can then be replaced by an integral) to get the final approximation. Since the indexing set is two dimensional, the linear order we use is the lexicographic ordering. So $(i_0, j_0)$ is the last time $Z_{ij}$ is greater than $b$ means that, $Z_{ij}$ is less than $b$ for all $(i, j) \in D_{i_0,j_0}$, where $D_{i_0,j_0} = \{(i, j) : j = j_0 \text{ and } i_0 < i < j_0 \text{ or } j > j_0 \text{ and } 0 < i < j \}$. So we get

$$P \left\{ \max_{1 \leq i < j \leq n} Z_{ij} > b \right\} = \sum_{1 \leq i_0 < j_0 \leq n} P \{ Z_{i_0,j_0} > b \text{ and } Z_{ij} \leq b, \text{ for all } (i, j) \in D_{i_0,j_0} \}.$$  \hspace{1cm} (A.10)

The probability on the right hand side can be written as follows.

$$P \{ Z_{i_0,j_0} > b \text{ and } Z_{ij} \leq b, \text{ for all } (i, j) \in D_{i_0,j_0} \}
= \int_b^\infty P \{ Z_{i_0,j_0} \in dx \} P \left\{ \max_{(i,j) \in D_{i_0,j_0}} Z_{ij} \leq b | Z_{i_0,j_0} = x \right\}
= \int_0^\infty \phi \left( b + \frac{x}{\sqrt{n}} \right) P \left\{ \max_{(i,j) \in D_{i_0,j_0}} \sqrt{n}(Z_{ij} - Z_{i_0,j_0}) \leq -x | Z_{i_0,j_0} = b + \frac{x}{\sqrt{n}} \right\} \frac{dx}{\sqrt{n}}, \hspace{1cm} (*)$$

where $\phi$ is the standard normal density.

In order to find the limiting value of the probability in $(*)$ we want the law of the $\sqrt{n}(Z_{ij} - Z_{i_0,j_0})$ conditioned on the event $Z_{i_0,j_0} = b + x n^{-1/2}$. Let $\delta_1, \delta_2, \delta_3$ be positive constants and let $D = \{(i_0,j_0) : i_0 > n \delta_1, \ j_0 < n (1 - \delta_2) \text{ and } j_0 - i_0 > n \delta_3 \}$. The calculation of the mean and the covariance function, similar to the one in (A.6) and
(A.7) shows that for \((i_0, j_0) \in D\), the law of the random variables in \((*)\) converges to the law of the random field

\[ U_k + V_l, \quad k = 0, \pm 1, \pm 2, \ldots; \quad l = 0, 1, 2, \ldots, \]

where \(U_k\) and \(V_l\) are mutually independent random walks with \(U_0 = V_0 = 0\) and increments \(N(\mu, \sigma^2)\) and \(N(\eta, \tau^2)\) respectively. Let \(s = i/n\), \(t = j/n\) and \(b_0 = b/\sqrt{n}\). Then \(\mu = b_0 t/[2s(t-s)], \sigma^2 = t/[s(t-s)], \eta = b_0 s/[2t(t-s)], \) and \(\tau^2 = s/[2t(t-s)]\).

Now, by using the random walk and renewal theory results in Siegmund (1989, Section 4), we get that

\[
\int_0^\infty \exp(-2bx/\sqrt{n})P \{(Z_{ij} - Z_{i_0j_0}) \leq -x \text{ for all } (i, j) \in D_{i_0j_0}|Z_{i_0j_0} = b + x\} \, dx
\]

converges to \(2\alpha \mu \eta \nu (2\mu/\sigma)\nu (2\eta/\tau)\), where \(\nu\) is as defined in (A.2) and \(\alpha = \mu/\sigma^2 = \eta/\tau^2\). So for \((i_0, j_0) \in D\) the probability in the sum in (A.10) converges to

\[
\frac{b^3 \phi(b)}{4} \int \frac{1}{(t-s)^2} \nu \left( b_0 \sqrt{\frac{t}{s(t-s)}} \right) \nu \left( b_0 \sqrt{\frac{s}{t(t-s)}} \right), \tag{A.11}
\]

where \(\nu\) is given in (A.2) and \(\phi\) is the standard normal density. Observe that, if we substitute (A.11) in (A.10) we get a Riemann sum which can be replaced by the limiting integral. Thus

\[
P \left\{ \max_{(i,j) \in D} Z_{ij} > b \right\} \approx \frac{b^3 \phi(b)}{4} \int \int \frac{1}{(t-s)^2} \nu \left( b_0 \sqrt{\frac{t}{s(t-s)}} \right) \nu \left( b_0 \sqrt{\frac{s}{t(t-s)}} \right) \, ds \, dt, \tag{A.12}
\]

where the integral is taken over the set \(\{(s, t) : s > \delta_1, t < 1 - \delta_2, t - s > \delta_3\}\). As in the binary segmentation case, we can set \(\delta_1 = \delta_2 = \delta_3 = 1/n\) to get the tail probability approximation for the likelihood ratio statistic. A simple change of variable will give us the formula in (A.4).

In Table A.1 we give the actual probability obtained by Monte Carlo simulations and the approximation for \(n = 200\) and a range of \(b\)'s. The simulated probability was obtained as follows. For each \(b\), we generated 10,000 random samples of size 200 each
and computed the maximum of the random field. The probability is the proportion of times the maximum was larger than the given value of $b$. The approximation is from the formula we derived.

Observe that the approximations are good near the tail of the distribution which is what is expected. Some corrections can be made to make the formula better.
Bibliography


