AN EXTENSION OF THE COMPOUND POISSON INFORMATION CRITERION TO 
AT MOST TWO CHANGE POINTS

by

Jacob S. Roach

An Abstract
presented in partial fulfillment
of the requirements for the degree of
Master of Science
in the Department of Mathematics and Computer Science
University of Central Missouri

May, 2014
In a data set, there could exist more than one set of parameters. Initially, the observed data could follow a distribution with one set of parameters and at a certain point the parameters of the distribution change. The data points where these changes occur are known as change points. In this thesis we examine the problem of locating change points given observations drawn from randomly generated data as part of a simulation study as well as gene expression data in application. The distribution of interest is a compound Poisson process which is the sum of normal observations. This paper uses an information criterion approach to identify at most two change points. Through simulation and application, this particular problem is examined.
AN EXTENSION OF THE COMPOUND POISSON INFORMATION CRITERION TO AT MOST TWO CHANGE POINTS

by

Jacob S. Roach

A Thesis
presented in partial fulfillment
of the requirements for the degree of
Master of Science
in the Department of Mathematics and Computer Science
University of Central Missouri

May, 2014
AN EXTENSION OF THE COMPOUND POISSON INFORMATION CRITERION TO
AT MOST TWO CHANGE POINTS

by

Jacob S. Roach

APPROVED:

[Signatures]

Thesis Chair

Thesis Committee Member

Thesis Committee Member

ACCEPTED:

[Signatures]

Chair, Department of Mathematics

and Computer Science

Dean, Graduate School

UNIVERSITY OF CENTRAL MISSOURI

WARRENSBURG, MISSOURI
ACKNOWLEDGMENTS

There are so many people I would like to thank for their support and patience. Everyone in my life is so wonderful and I am so grateful for them. Foremost, I would like to thank my wife, my best friend, Jessica, for her unwavering support of my decisions and extraordinary patience as we have waited to begin our lives together. My mother, father, and grandmother raised me to never give up and I hope they can see their reflections in me. I would like to thank the faculty and staff for their hard work; especially my advisor, Dr. Plummer. Over the past two years, I have learned so much from my students, and I thank all of them for being such outstanding individuals. The application of this paper is to cancer research. During the end of writing this paper, I drew strength and inspiration from individuals and family members of individuals affected by the disease. I pray that something I have done will help someone become well. As with everything in my life, this work is dedicated to my Lord and Savior. He guides my footsteps as I walk the path he has chosen for me.
Contents

List of Figures viii

List of Tables ix

1 Preliminaries 1

1.1 Compound Poisson Process . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
1.2 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
1.3 Outline of Remaining Chapters . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2 Introduction To Change Point Analysis and Review of Literature 4

2.1 Introduction to Change Point Analysis . . . . . . . . . . . . . . . . . . . . . . . 4
2.2 Parametric Methods in Fixed Set Change Point Analysis . . . . . . . . . . . . 5
2.3 Other Well Studied Distributions in Statistical Change Point Analysis . . . 7

3 Statement of the Problem 8

3.1 Generalized Statement of The Change Point Problem . . . . . . . . . . . . . 8
3.2 Statement of the Change Point Problem Studied . . . . . . . . . . . . . . . . . 9
3.3 Preface of the Simulation Study . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

4 Change Points in a Compound Poisson Process with Variance Known 13

4.1 Overview of the Information Criterion Approach . . . . . . . . . . . . . . . . . 13
# CONTENTS

4.2 The Likelihood Function and MLEs for the Case of at Most One Change Point 15
   4.2.1 Development of the MLEs Under $H_0$ .......................... 15
   4.2.2 Development of the MLEs Under $H_1$ .......................... 16

4.3 The Likelihood Function and MLEs for the Case of Two Change Points . . 18

4.4 Compound Poisson Information Criterion .............................. 21

5 Simulation Study and Application to aCGH Data 26
   5.1 Overview ................................................................. 26
   5.2 Simulation Study ....................................................... 27
      5.2.1 Simulations with Two Change Points ......................... 27
      5.2.2 Simulations with Zero Change Points ....................... 34
      5.2.3 Simulations with a Single Change Point .................... 36
   5.3 Application to aCGH Data Set .......................... 37

A R Code 41
   A.1 Simulation R Code ........................................ 41
   A.2 Application R Code ........................................ 47
   A.3 Chromosome Plots ........................................ 51

Bibliography 53
List of Figures

5.1 Simulated Compound Poisson Data for N=200 .......................... 28
5.2 Exponential Variable for Zero Change Point Simulation with False Positive .......................... 36
5.3 Normal Variable for Zero Change Point Simulation with False Positive .......................... 36
5.4 Chromosome with Two Change Points Correctly Identified .......................... 38
5.5 Single Change Point Detected, Two Change Point Model Selected .......................... 39
5.6 False Positives From Outlier .......................... 39
5.7 Chromosome with a Single Change Point Detected Unaffected by Outlier .......................... 40
A.1 GM01535 Chromosome 5 ......................................................... 51
A.2 GM01535 Chromosome 12 ......................................................... 51
A.3 GM01750 Chromosome 14 ......................................................... 51
A.4 GM13330 Chromosome 1 ......................................................... 52
A.5 GM13330 Chromosome 4 ......................................................... 52
A.6 GM13031 Chromosome 17 ......................................................... 52
List of Tables

5.1 Results Table for \((\mu_1, \lambda_1) = (2, 2)\) ................................................. 29
5.2 Results Table for \((\mu_1, \lambda_1) = (3, 3)\) ................................................. 30
5.3 Results Table for \((\mu_1, \lambda_1) = (3.5, 4)\) ................................................. 31
5.4 Frequency of Selecting Two Change Point Model and Positions .................. 32
5.5 Accuracy of Model Selection for Second Change Point at Location \(j_1 + 5\) ........... 33
5.6 Null Simulation Results ........................................................................ 35
5.7 aCGH Application Results Table .............................................................. 38
Chapter 1

Preliminaries

1.1 Compound Poisson Process

In this section, we will introduce the definition of a Compound Poisson Process.

Suppose the probability of some event that occurs follows POI(\(\lambda\)). When the event occurs, a measurement \(Y_i\) is taken and \(Y_i\) for \(i = 1, 2 \ldots, N\) are independently and identically distributed random variables. In an interval \([0, T]\) we have \(N\) occurrences. Thus, \(N \sim \text{POI}(\lambda T)\).

Partition \([0, T]\) into sub-intervals of length \(T_i, i = 1, 2, \ldots, m\). In each sub-interval, let \(n_i\) be the number of occurrences of the event. We have that \(n_i \sim \text{POI}(\lambda T_i)\). Let \(t_i\) denote the \(i^{th}\) sub-interval of \([0, T]\). Let \(X_i = \sum_{Y_j \in t_i} Y_j\). Then \(X_i\) is a compound Poisson variable and \(X_1, X_2, \ldots, X_m\) form a compound Poisson process. Note that the \(X_i\)'s are independent but not identically distributed as \(n_i\) are not assumed to be equal, for all \(i\).
CHAPTER 1. PRELIMINARIES

1.2 Motivation

Compound Poisson processes in the area of statistical change point analysis compose an understudied problem. The review of the relevant literature of the subject will reflect this notion. The academic value of the study of compound Poisson processes in statistical change point analysis is not limited to statistical theory. Change point analysis has applications in many different areas. The areas of application include, but are not limited to, Stock Market Analysis, Quality Control in manufacturing, Geology, Psychology, and Finance [2].

The study of compound Poisson processes in change point analysis is also applied to gene expression data. Plummer and Chen studied DNA copy number variations using a Bayesian approach for locating change points in a compound Poisson process [15]. This study yields a comparison between copy number variations in normal cells versus cancer cells. Mutations in certain genes can be lack of production, known as deletions, or increased production, known as amplifications of DNA sequences. Locating the deletions or amplifications is useful in developing medical treatments [15]. The application to gene expression data is valuable to the study of compound Poisson processes in change point analysis.

In this paper, the method developed will be tested in simulations and with gene expression data. The gene expression data used in this thesis are useful in determining the ability of the method used to identify change points in a compound Poisson process as the location(s) of the change point(s) are known and have been verified through biological methods. Applying the results of the research to simulated and actual data sets will assist in determining the effectiveness of the method developed in locating change points in a compound Poisson process as defined later in this work.
1.3 Outline of Remaining Chapters

The following chapters are structured as follows. In the next chapter, the field of change point analysis will be formally introduced. A review of previous works and applications in the field will be discussed; in which we will highlight that the multiple change point problem is not as heavily studied as the single change point problem. The review will also emphasize that compound Poisson processes in change point analysis is also understudied. For these reasons, this paper addresses the multiple change point problem under the assumption of a compound Poisson process.

The statement of the problem to be studied in this paper is given in Chapter 3. Chapter 3 will also introduce the approach taken to study the problem. Chapter 4 will give an overview of the information criterion approach. Later in Chapter 4, Compound Poisson Information Criterion, known as CPIC, will be given. As part of the information criterion approach, CPIC is used in hypothesis testing. In order to utilize the information criterion approach it is necessary to show the development of the likelihood function based on the probability density function coinciding with the compound Poisson process. Both the log likelihood function and maximum likelihood estimators will be derived and the CPIC penalty term will be specified. The statistics derived in chapter 4 will be applied to build a proper hypothesis test in accordance to the information criterion approach.

Chapter 5 presents the simulation study performed and the results from the simulations are explored. Also included in this chapter is the application of the CPIC approach to sets of aCGH (Array Comparative Genomic Hybridization) data to further show the effectiveness of the method developed.
Chapter 2

Introduction To Change Point Analysis and Review of Literature

2.1 Introduction to Change Point Analysis

Compared to many other areas of statistics, change point analysis is relatively modern. E.S. Page published papers in 1955 and 1957 that formalized the field of change point analysis. Since Page’s works in the 1950’s, the field of change point analysis has grown into many different areas. Some areas have been more heavily studied than others. Topics where at most a single change exists have been studied more so than the case of multiple change points[2].

Change point analysis can be divided into two distinct areas: fixed set, also known as off-line change point analysis, and sequential or on-line change point analysis. In fixed set change point analysis pre-observed data are given in the order of observation. In sequential change point analysis repeated measurements are taken as part of an ongoing process. As well as locating the change, sequential change point analysis emphasizes identifying the change as soon as possible after the change occurs [14]. Both fixed set and sequential change point
analysis have been studied extensively using both parametric and non-parametric statistical methods. The focus of this paper is a compound Poisson process in fixed set change point analysis using an information criterion approach, which falls into the category of parametric statistics. Therefore, this review of literature is limited to the area of parametric statistical change point analysis in the area of fixed set.

For the purposes of this work, a change point can be defined as a place or time point such that the observations follow a given distribution up to the change point and follow a different distribution after the change point. This describes the case of a single change point. For $n$ number of change points, the distribution would change $n$ times. Of primary concern is the detection of the location of the change point(s). A challenge in the case of multiple change point models is the determination of the number of changes which have occurred in addition to their location. Thus, we can view change point analysis as a process in which to decide if the observed data follows a common distribution throughout the sequence versus a change in distribution for at least one point [12].

2.2 Parametric Methods in Fixed Set Change Point Analysis

In fixed set change point analysis the data is pre-observed and given in the order of observation. The problem to be studied involves identifying the location at which the parameters change [14].

Note: Due to the complexity of the multiple change point problem, in the following paragraphs, the following works discussed are applied to the single change point; unless stated otherwise.

There have been many approaches taken to address the single change point problem in a
set of normal observations. During the emergence of the field, Page (1957) used a CUSUM method [12]. Chernoff and Zacks (1969) used a Bayesian approach [3]. Hinkley (1970) derived the asymptotic distribution for the maximum likelihood estimate [7]. Hawkins (1977) derived the distribution of the Likelihood Ratio Test Statistic along with its asymptotic distribution [6]. All of these methods focused on a change in mean with variance known. Hawkins extended this work and included results for the case of a fixed, unknown variance [6]. However, Worsley (1979) corrected Hawkins’ work for the case of unknown variance [17].

In the past few decades, the problem of change in mean has resurfaced. Zhang and Siegmund (2007) developed a Modified Bayesian Information Criterion [18]. This method is known as modified BIC [18]. Modified BIC was applied to detect the number and location(s) of the change point(s). The modified BIC approach allowed for the study of more than a single change point. Erdman and Emerson (2008) also studied the multiple change point problem, proposing a Bayesian method [4]. Finally, Vositrikova (1981) used a process called binary segmentation to approach the multiple change point problem, and proved its consistency. This process searches for a single change point in the data set. If a change point is found and determined to be significant, the data set is segmented at that location. The process is repeated for each segment until no new change points are detected [16].

Another problem of interest is to assume a fixed mean and find the location of a shift in variance. Hsu (1977) studied this problem under the assumption that the initial value of the variance was known. Hsu developed two tests; one of which was based on the CUSUMs of the Chi Squared Values [8]. A Bayesian approach to identify multiple change points was taken by Inclan (1993) [9]. Chen and Gupta (1997) used the Schwarz information criterion to identify a single change point assuming a fixed mean [1].
CHAPTER 2. INTRODUCTION TO CHANGE POINT ANALYSIS AND REVIEW OF LITERATURE

2.3 Other Well Studied Distributions in Statistical Change Point Analysis

In *Parametric Statistical Change Point Analysis With Applications to Genetics, Medicine, and Finance*, Chen and Gupta discuss other sub-topics in the field [2]. They include sections on the following distributions: Gamma, Exponential, Binomial, Poisson, and Multivariate Normal. Modified Information Criterion was developed for the exponential family of distributions. A change in distribution is not the only topic of interest in change point analysis. Another is the occurrence in change of statistical model; such as regression models, time series models, and Hazard Functions. Chen and Gupta include these topics, as well. It is noteworthy to say that these topics are well studied [2].

Since the formalization of change point analysis in the 1950s, many topics have been well studied. Locating change points in a compound Poisson process, however, is not a well studied topic. Plummer and Chen published an article in 2013 for locating change points in a compound Poisson process using a Bayesian approach. Their work included application to detecting DNA copy number variations [15]. Ng (2008) used an EM algorithm to estimate the maximum likelihood estimators’ efficiency [10]. Ng’s method was then demonstrated using a compound Poisson process, where the variables being summed were exponentially distributed. Most of the works in the area of change point analysis are in areas other than the case of a compound Poisson process. For this reason, the focus and motivation for the research presented in this paper is devoted to the case of a compound Poisson process in fixed set change point analysis extended to multiple change points. As a simplifying restriction, we assume at most two change points.
Chapter 3

Statement of the Problem

3.1 Generalized Statement of The Change Point Problem

Note: The generalized statement of the problem is in accordance to the motivation presented in section 1.2.

Statistical change point analysis can be viewed as a decision making process to determine if the observations follow a common distribution. From this standpoint, the change point problem is a hypothesis testing problem.

E.S. Page described the change point problem from a hypothesis testing perspective. In “A Test for a Change in a Parameter Occurring at an Unknown Point”, Page describes the problem [11].

A sample of independent observations are obtained, \(x_1, x_2, \ldots, x_n\). To test the null hypothesis that all the observations are drawn from the same population with distribution function \(F(x|\theta)\) against the alternative that the observations, \(x_1, x_2, \ldots, x_m\) come from \(F(x|\theta)\) and \(x_{m+1}, x_{m+2}, \ldots, x_n\) come from \(F(x|\theta')\) where \(\theta \neq \theta'\) [11].
Page’s statement of the problem is similar to the statement of the problem in this paper. However, in this study we extend beyond a single change point, and two sets of parameters, to investigate cases where the observations have at most two change points. Thus, the observations could have at most three sets of parameters.

It is noteworthy to mention that this statement can be extended further to $p$ change points, $p \in \mathbb{N}$.

### 3.2 Statement of the Change Point Problem Studied

The statement of the problem to be examined in this paper is as follows:

Suppose an event occurs according to a Poisson process and that when the event occurs a measurement, $Y_i$, is taken. Assume $Y_i \sim iid \ N(\mu, \sigma^2)$ for $i = 1, 2, \ldots, M$; where $M$ is the number of occurrences of the event over a time or distance $T$. Then $M \sim POI(\lambda T)$; where $\lambda$ is the unit rate.

Let $[0,T]$ be partitioned into $l$ sub-intervals of length $T_j$, $j = 1, 2, \ldots, l$. Then there exist $m_j$ occurrences in each $T_j$ and $m_j \sim POI(\lambda T_j)$. Let $M_j = \sum_{i=1}^{J} m_i$ with $M_0 = 0$. For $j = 1, 2, \ldots, l$, we define $X_{t_j} = \sum_{i=M_{j-1}+1}^{M_j} Y_i$, where $t_j$ denotes the $j^{th}$ sub-interval on $[0,T]$. Then, $X_{t_1}, X_{t_2}, \ldots, X_{t_l}$ forms a compound Poisson process dependent on $(\mu, \lambda, \sigma^2)$. Given $m_j$, $X_{t_j} \sim N(m_j \mu, m_j \sigma^2)$ and $X_{t_1}, X_{t_2}, \ldots, X_{t_l}$ are independent.
Provided the sequence of observations, as stated above, the change point problem, for any number of change points, can be tested using the following hypothesis test:

\[ H_0 : \quad X_{t_1}, X_{t_2}, \ldots, X_{t_l} \text{ form a compound Poisson process} \]

depending on parameters \( \theta \) where, \( \theta \) is the dimension of the model, 
\[
\theta = (\mu_0, \lambda_0, \sigma^2) \quad \text{and for any } \theta' \neq \theta \]

versus,

\[ H_1 : \quad X_{t_1}, X_{t_2}, \ldots, X_{t_{j_1}} \text{ depend on } \theta' = (\mu_1, \lambda_1, \sigma^2), \]
\[
X_{t_{j_1}+1}, X_{t_{j_1}+2}, \ldots, X_{t_{j_2}} \text{ depend on } \theta'' = (\mu_2, \lambda_2, \sigma^2), \]
\[
X_{t_{j_2}+1}, X_{t_{j_2}+2}, \ldots, X_{t_{j_3}} \text{ depend on } \theta''' = (\mu_3, \lambda_3, \sigma^2), \ldots, \]
\[
X_{t_{j_p+1}}, X_{t_{j_p+2}}, \ldots, X_{t_l} \text{ depend on } \theta^{(p)} = (\lambda_{p+1}, \mu_{p+1}, \sigma^2), \text{ where,} \]
\[
\theta' = (\lambda_1, \mu_1, \sigma^2) \neq \theta'' = (\lambda_2, \mu_2, \sigma^2) \neq \theta''' = (\lambda_3, \mu_3, \sigma^2) \neq \]
\[
\ldots \neq \theta^{(p+1)} = (\mu_{p+1}, \lambda_{p+1}, \sigma^2) \]

For \( p \) number of change points, \( p \in \mathbb{N} \).

Due to the complexity of the general case, this paper will focus on the case of at most two change points. The case of at most two change points can be stated using the following nested hypotheses:
CHAPTER 3. STATEMENT OF THE PROBLEM

\( H_0 \) : \( X_{t_1}, X_{t_2}, \ldots, X_{t_\theta} \) form a compound Poisson process depending on parameters \( \theta \) where, \( \theta \) is the dimension of the model,

\[ \theta = (\mu_0, \lambda_0, \sigma^2) \] and for any \( \theta', \theta = \theta' \) \( (3.3) \)

versus,

\( H_1 \) : \( X_{t_1}, X_{t_2}, \ldots, X_{t_{j_1}} \) depend on \( \theta' = (\lambda_1, \mu_1, \sigma^2) \), \( X_{t_{j_1}+1}, X_{t_{j_1}+2}, \ldots, X_{t_{j_2}} \) depend on \( \theta'' = (\lambda_2, \mu_2, \sigma^2) \), where \( \theta' = (\lambda_1, \mu_1, \sigma^2) \neq \theta'' = (\lambda_2, \mu_2, \sigma^2) \) \( (3.4) \)

and,

\( H_0 \) : \( X_{t_1}, X_{t_2}, \ldots, X_{t_{j_1}} \) depend on \( \theta' = (\lambda_1, \mu_1, \sigma^2) \), \( X_{t_{j_1}+1}, X_{t_{j_1}+2}, \ldots, X_{t_{j_2}} \) depend on \( \theta'' = (\lambda_2, \mu_2, \sigma^2) \), where \( \theta' = (\lambda_1, \mu_1, \sigma^2) \neq \theta'' = (\lambda_2, \mu_2, \sigma^2) \) \( (3.5) \)

versus,

\( H_1 \) : \( X_{t_1}, X_{t_2}, \ldots, X_{t_{j_1}} \) depend on \( \theta' = (\lambda_1, \mu_1, \sigma^2) \), \( X_{t_{j_1}+1}, X_{t_{j_1}+2}, \ldots, X_{t_{j_2}} \) depend on \( \theta'' = (\lambda_2, \mu_2, \sigma^2) \), \( X_{t_{j_2}+1}, X_{t_{j_2}+2}, \ldots, X_{t_{j_3}} \) depend on \( \theta''' = (\lambda_3, \mu_3, \sigma^2) \), where \( \theta' = (\lambda_1, \mu_1, \sigma^2) \neq \theta'' = (\lambda_2, \mu_2, \sigma^2) \neq \theta''' = (\lambda_3, \mu_3, \sigma^2) \) \( (3.6) \)
3.3 Preface of the Simulation Study

The statistics presented in Chapter 4 was applied to create a program designed for R Statistical Analysis Software. The program generated data sets based on user defined parameters. A specialized algorithm was designed to identify zero, one, or two change points within the generated data. The algorithm would select the most appropriate model for the data. The results of the simulations were tabulated for the purpose of reviewing the performance of the algorithm.

Simulation of this design is used as a bench mark to evaluate the correctness of the mathematics and statistics involved in the research of a compound Poisson process in change point analysis.

Once the performance of the algorithm was validated, the process was applied to gene expression data, known as aCGH (Array Comparative Genome Hybridization) data. This application was for the purpose of studying the process developed in the research against real world data. Test data sets have known location(s) of change point(s) verified using biological methods.

There exist inherent differences in the types of data in the randomly generated data and the aCGH data. These differences are detailed in Chapter 5. Therefore, additional work was directed towards adapting the algorithm to account for these differences.

The goal of such extensive simulation and application was to examine the performance of the algorithm, thus, enhancing the study of the problem given in this paper according to the motivation.
Chapter 4

Change Points in a Compound
Poisson Process with Variance Known

4.1 Overview of the Information Criterion Approach

For the particular change point problem studied in this paper, we will assume a known constant variance, $\sigma^2$, with a possible change in the mean, $\mu$, and rate of occurrence, $\lambda$. It has been shown that the Maximum Likelihood Estimators, MLEs, for this case are proven consistent estimators, while the same has yet to be proved about the cases involving unknown variance [14]. Therefore, this paper will not discuss the case with variance unknown.

The change point problem with known variance can be approached using information criterion. This approach is advantageous to this hypothesis test as the distribution of the likelihood ratio procedure test is as yet unknown [14]. Information criterion is used to select between competing statistical models [13]. Information criterion, such as the Akaike information criterion and the Schwarz information criterion, are designed to select models with simple structure and interpretive value [5]. The complexity is typically measured by the number of parameters being estimated in the model [14]. The number of parameters
of the model is the dimension, denoted $\text{dim}(\theta)$, as previously mentioned in the statement of the problem. A penalty term is used to compensate for models that could be a best fit model, but not be the most appropriate model for the application. In this paper, the penalty term is denoted by $\kappa$. Conceptually, models with additional, unnecessary, parameters are penalized, thus guaranteeing the negative two log maximum likelihood function, $-2\log L(\hat{\theta})$, must improve more than the additional penalty for a more complex model to be selected. The penalty term, $\kappa$, reduces the rate of false positive selections [14]. For the purpose of the study of the compound Poisson model, the information criterion used is Compound Poisson Information Criterion; referred to as CPIC. CPIC was developed by Plummer as a new modified information criterion [14].

The goal of the information criterion approach is to develop a calculation procedure to test for change points in application. As part of this development, the likelihood function for the null and alternative hypotheses will be presented. Then the log likelihood function will be derived. For each parameter, Maximum Likelihood Estimators, MLEs, must be calculated to allow for hypothesis testing. The MLEs will be used to estimate $\theta$. From this, we will present the generalized CPIC equation. The generalized CPIC equation is used in practical application to test for change points.

The remaining sections in this chapter will show the development of the procedure as outlined above in accordance to the statement of the problem.
CHAPTER 4. CHANGE POINTS IN A COMPOUND POISSON PROCESS WITH VARIANCE KNOWN

4.2 The Likelihood Function and MLEs for the Case of at Most One Change Point

4.2.1 Development of the MLEs Under \( H_0 \)

First, we will examine the likelihood function and Maximum Likelihood Estimators under \( H_0 \) for at most one change point developed by Plummer [14].

Let \((X, N)\) form a compound Poisson process of normal variables with mean \( \mu \), variance \( \sigma^2 \), and \( \lambda \) the unit rate of the Poisson process with \( T \) the length of the interval measured.

The pdf is given as,

\[
f(x, n) = \begin{cases} 
\frac{1}{\sqrt{2\pi n\sigma^2}} e^{-\frac{(x-n\mu)^2}{2n\sigma^2}} (T\lambda)^n e^{-\frac{T\lambda}{(n!)}} : n \neq 0 \\
e^{-T\lambda} : n = 0
\end{cases}
\]  

For sample observations as described in the statement of the problem, \((x_{t1}, m_1), (x_{t2}, m_2), \ldots, (x_{tl}, m_l)\), let \( A = \{i | m_i \neq 0\}, |A| = M_A \). The likelihood function under \( H_0 \) is as follows:

\[
L_0(\mu, \lambda) = \left( \prod_{i \in A} \frac{1}{\sqrt{2\pi m_i\sigma^2}} e^{-\frac{(x_{ti}-m_i\mu)^2}{2m_i\sigma^2}} \right) \left( \prod_{i=1}^{l} \frac{(T_i\lambda)^{m_i} e^{-T_i\lambda}}{m_i!} \right) 
= (2\pi\sigma^2)^{-\frac{M_A}{2}} \left( \prod_{i \in A} m_i^{-1/2} \right) e^{-\frac{1}{2\sigma^2} \sum_{i \in A} \frac{(x_{ti}-m_i\mu)^2}{m_i}} \cdot \left( \prod_{i=1}^{l} \frac{T_i^{m_i}}{m_i!} \right) \sum_{i=1}^{l} m_i \cdot e^{-\lambda \sum_{i=1}^{l} T_i} 
\]  

(4.2)
Taking the log of the likelihood function under $H_0$, we obtain:

\[
\log(L_0) = -\frac{M_A}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{i \in A} \log(m_i) - \frac{1}{2\sigma^2} \sum_{i \in A} \frac{(x_t - m_i\mu)^2}{m_i} \\
+ \sum_{i=1}^l m_i \log(T_i) - \sum_{i=1}^l \log(m_i!) + \log(\lambda) \sum_{i=1}^l m_i - \lambda \sum_{i=1}^l T_i \tag{4.3}
\]

Plummer showed that the MLEs for $\mu$ and $\lambda$ as follows:

\[
\hat{\mu} = \frac{\sum_{i=1}^l x_t}{\sum_{i=1}^l m_i} \tag{4.4}
\]

\[
\hat{\lambda} = \frac{\sum_{i=1}^l m_i}{\sum_{i=1}^l T_i} \tag{4.5}
\]

Note that while the partial derivative with respect to $\mu$ contains $i \in A$, the MLEs can be written as the sums from 1 to $l$ since for $i \notin A$ both $x_i = 0$ and $m_i = 0$.

**4.2.2 Development of the MLEs Under $H_1$**

Now that the likelihood function and MLEs for $\mu$ and $\lambda$ have been calculated under $H_0$, we can examine the likelihood function and MLEs under $H_1$ for at most one change point.

Let

\[
A_j = \{i | m_i \neq 0, \ 1 \leq i \leq j\}, \text{ such that } |A_j| = M_{A_j},
\]

\[
A_l = \{i | m_i \neq 0, \ j + 1 \leq i \leq l\}, \text{ such that } |A_l| = M_{A_l}
\]
CHAPTER 4. CHANGE POINTS IN A COMPOUND POISSON PROCESS WITH VARIANCE KNOWN

Then, the likelihood function for at most a single change point is given as:

\[
L_1(\mu_1, \lambda_1, \mu_2, \lambda_2) = \left( \prod_{i \in A_j} \frac{1}{\sqrt{2\pi m_i \sigma^2}} e^{-\left(\frac{(x_{t_i} - m_i \mu_1)^2}{2m_i \sigma^2}\right)} \right) \left( \prod_{i=1}^{l} \frac{(T_i \lambda_1)^{m_i} e^{-T_i \lambda_1}}{m_i!} \right) \cdot \left( \prod_{i \in A_l} \frac{1}{\sqrt{2\pi m_i \sigma^2}} e^{-\left(\frac{(x_{t_i} - m_i \mu_2)^2}{2m_i \sigma^2}\right)} \right) \left( \prod_{i=j+1}^{l} \frac{(T_i \lambda_2)^{m_i} e^{-T_i \lambda_2}}{m_i!} \right) \\
= (2\pi \sigma^2)^{-M_A} \left( \prod_{i \in A_j} (m_i)^{-1/2} \right) e^{-\left(\frac{1}{2\sigma^2}\right) \sum_{i \in A_j} \frac{(x_{t_i} - m_i \mu_1)^2}{m_i} + \sum_{i \in A_l} \frac{(x_{t_i} - m_i \mu_2)^2}{m_i}} \cdot \left( \prod_{i=1}^{l} \frac{T_i^{m_i}}{m_i!} \right) \lambda_1^{\sum_{i=1}^{j} m_i} \lambda_2^{\sum_{i=j+1}^{l} m_i} e^{-\lambda_1 \sum_{i=1}^{j} T_i} - \lambda_2 \sum_{i=j+1}^{l} T_i \right) (4.6)
\]

The log likelihood function under \( H_1 \) for at most one change point is given as:

\[
\log L_1(\mu_1, \lambda_1, \mu_2, \lambda_2) = -\frac{M_A}{2} \log(2\pi \sigma^2) + \left( -\frac{1}{2} \right) \sum_{i \in A_j} \log(m_i) \\
- \left( \frac{1}{2\sigma^2} \right) \left[ \sum_{i \in A_j} \frac{(x_{t_i} - m_i \mu_1)^2}{m_i} + \sum_{i \in A_l} \frac{(x_{t_i} - m_i \mu_2)^2}{m_i} \right] \\
+ \sum_{i=1}^{l} m_i \log(T_i) - \sum_{i=1}^{l} \log(m_i) \\
+ \log(\lambda_1) \sum_{i=1}^{j} m_i + \log(\lambda_2) \sum_{i=j+1}^{l} m_i \\
- \lambda_1 \sum_{i=1}^{j} T_i - \lambda_2 \sum_{i=j+1}^{l} T_i \right) (4.7)
\]
Denote $\hat{\mu}_i, \hat{\lambda}_i$ as the MLEs for $\mu_i, \lambda_i$

Plummer showed that the MLEs for $\mu_i, \lambda_i$, $i \in \{1, 2\}$ are as follows [14]:

\[
\hat{\mu}_1 = \frac{\sum_{i=1}^{j} x_{t_i}}{\sum_{i=1}^{j} m_i} \quad (4.8)
\]

\[
\hat{\mu}_2 = \frac{\sum_{i=j+1}^{l} x_{t_i}}{\sum_{i=j+1}^{l} m_i} \quad (4.9)
\]

\[
\hat{\lambda}_1 = \frac{\sum_{i=1}^{j} m_i}{\sum_{i=1}^{j} T_i} \quad (4.10)
\]

\[
\hat{\lambda}_2 = \frac{\sum_{i=j+1}^{l} m_i}{\sum_{i=j+1}^{l} T_i} \quad (4.11)
\]

4.3 The Likelihood Function and MLEs for the Case of Two Change Points

Now we will develop the likelihood function under $H_1$ for at most two change points. For at most two change points, the observations must be partitioned into three subsets, in which each partition occurs at the location of a change point.

Let $A_{j_1} = \{i|m_i \neq 0, \ 1 \leq i \leq j_1\}$, such that $|A_{j_1}| = M_{A_{j_1}}$

$A_{j_2} = \{i|m_i \neq 0, \ j_1 + 1 \leq i \leq j_2\}$, such that $|A_{j_2}| = M_{A_{j_2}}$

$A_{l} = \{i|m_i \neq 0, \ j_2 + 1 \leq i \leq l\}$, such that $|A_{l}| = M_{A_{l}}$

The likelihood function under $H_1$ for at most two change points is as follows:
CHAPTER 4. CHANGE POINTS IN A COMPOUND POISSON PROCESS WITH VARIANCE KNOWN

\[ L_1(\mu_1, \lambda_1, \mu_2, \lambda_2, \mu_3, \lambda_3) = \left( \prod_{i \in A_{j_1}} \frac{1}{\sqrt{2\pi m_i \sigma^2}} e^{-\frac{(x_{t_i} - m_i \mu_1)^2}{2m_i \sigma^2}} \right) \left( \prod_{i=1}^{j_1} (T_i \lambda_1)^{m_i} e^{-T_i \lambda_1} \right) \]

\[ \cdot \left( \prod_{i \in A_{j_2}} \frac{1}{\sqrt{2\pi m_i \sigma^2}} e^{-\frac{(x_{t_i} - m_i \mu_2)^2}{2m_i \sigma^2}} \right) \left( \prod_{i=j_1+1}^{j_2} (T_i \lambda_2)^{m_i} e^{-T_i \lambda_2} \right) \]

\[ \cdot \left( \prod_{i \in A_{j_3}} \frac{1}{\sqrt{2\pi m_i \sigma^2}} e^{-\frac{(x_{t_i} - m_i \mu_3)^2}{2m_i \sigma^2}} \right) \left( \prod_{i=j_2+1}^{j_3} (T_i \lambda_3)^{m_i} e^{-T_i \lambda_3} \right) \]

\[ = \left( 2\pi \sigma^2 \right)^{-\frac{M_A}{2}} \left( \prod_{i \in A} (m_i)^{-1/2} \right) e^{-\frac{1}{2\sigma^2} \left[ \sum_{i \in A_{j_1}} \frac{(x_{t_i} - m_i \mu_1)^2}{m_i} + \sum_{i \in A_{j_2}} \frac{(x_{t_i} - m_i \mu_2)^2}{m_i} + \sum_{i \in A_{j_3}} \frac{(x_{t_i} - m_i \mu_3)^2}{m_i} \right]} \]

\[ \cdot \left( \prod_{i=1}^{l} \frac{T_i^{m_i}}{m_i!} \right) \frac{1}{\lambda_1} \frac{1}{\lambda_2} \frac{1}{\lambda_3} e^{-\lambda_1 \sum_{i=1}^{j_1} T_i - \lambda_2 \sum_{i=j_1+1}^{j_2} T_i - \lambda_3 \sum_{i=j_2+1}^{j_3} T_i} \]

(4.12)

From the likelihood function above, the log likelihood function under \( H_1 \), \( \log L_1(\theta) \), where,

\[ L_1(\mu_1, \lambda_1, \mu_2, \lambda_2, \mu_3, \lambda_3) = L_1(\theta) \] is as follows:
\[ \log L_1(\theta) = -\frac{M_A}{2} \log(2\pi\sigma^2) + \left( -\frac{1}{2} \right) \sum_{i \in A} \log(m_i) \]

\[ - \left( \frac{1}{2\sigma^2} \right) \left[ \sum_{i \in A_{j_1}} (x_{i} - m_i\mu_1)^2 + \sum_{i \in A_{j_2}} (x_{i} - m_i\mu_2)^2 + \sum_{i \in A_{l}} (x_{i} - m_i\mu_3)^2 \right] \]

\[ + \sum_{i=1}^{l} m_i \log(T_i) - \sum_{i=1}^{l} \log(m_i!) + \sum_{i=1}^{j_1} m_i \log(\lambda_1) + \sum_{i=j_1+1}^{j_2} m_i \log(\lambda_2) \]

\[ + \sum_{i=j_2+1}^{l} m_i \log(\lambda_3) - \lambda_1 \sum_{i=1}^{j_1} T_i - \lambda_2 \sum_{i=j_1+1}^{j_2} T_i - \lambda_3 \sum_{i=j_2+1}^{l} T_i \] (4.13)

As under \( H_0 \), a maximum for the estimation of the parameters under \( H_1 \) can be found by setting the partial derivative of the log likelihood function with respect to each parameter equal to zero and solving [14]. Denote \( l_1 = \log L_1(\hat{\theta}) \)

The following system of partials are derived:

\[ \frac{\partial l_1}{\partial \mu_1} = \frac{1}{\sigma^2} \left[ \sum_{i \in A_{j_1}} x_{i} - \mu_1 \sum_{i \in A_{j_1}} m_i \right] \] (4.14)

\[ \frac{\partial l_1}{\partial \mu_2} = \frac{1}{\sigma^2} \left[ \sum_{i \in A_{j_2}} x_{i} - \mu_2 \sum_{i \in A_{j_2}} m_i \right] \] (4.15)

\[ \frac{\partial l_1}{\partial \mu_3} = \frac{1}{\sigma^2} \left[ \sum_{i \in A_{l}} x_{i} - \mu_3 \sum_{i \in A_{l}} m_i \right] \] (4.16)

\[ \frac{\partial l_1}{\partial \lambda_1} = \frac{1}{\lambda_1} \sum_{i=1}^{j_1} m_i - \sum_{i=1}^{j_1} T_i \] (4.17)
CHAPTER 4. CHANGE POINTS IN A COMPOUND POISSON PROCESS WITH VARIANCE KNOWN

\[
\frac{\partial l_1}{\partial \lambda_2} = \frac{1}{\lambda_2} \sum_{i=j_1+1}^{j_2} m_i - \sum_{i=j_1+1}^{j_2} T_i \tag{4.18}
\]

\[
\frac{\partial l_1}{\partial \lambda_3} = \frac{1}{\lambda_3} \sum_{i=j_2}^{l} m_i - \sum_{i=j_2+1}^{l} T_i \tag{4.19}
\]

Setting \(\frac{\partial l_1}{\partial \mu_i} = 0\) and \(\frac{\partial l_1}{\partial \lambda_i} = 0\) we can solve for the MLEs for each \(\mu_i\) and \(\lambda_i, i \in \{1, 2, 3\}\).

Denoting the MLEs for \(\mu_i, \lambda_i\) as \(\hat{\mu}_i, \hat{\lambda}_i\), the MLEs for \(\mu_i, \lambda_i, i \in \{1, 2, 3\}\) are as follows:

\[
\hat{\mu}_1 = \frac{\sum_{i=1}^{j_1} x_{ti}}{\sum_{i=1}^{j_1} m_i} \tag{4.20}
\]

\[
\hat{\mu}_2 = \frac{\sum_{i=j_1+1}^{j_2} x_{ti}}{\sum_{i=j_1+1}^{j_2} m_i} \tag{4.21}
\]

\[
\hat{\mu}_3 = \frac{\sum_{i=j_2+1}^{l} x_{ti}}{\sum_{i=j_2+1}^{l} m_i} \tag{4.22}
\]

\[
\hat{\lambda}_1 = \frac{\sum_{i=1}^{j_1} m_i}{\sum_{i=1}^{j_1} T_i} \tag{4.23}
\]

\[
\hat{\lambda}_2 = \frac{\sum_{i=j_1+1}^{j_2} m_i}{\sum_{i=j_1+1}^{j_2} T_i} \tag{4.24}
\]

\[
\hat{\lambda}_3 = \frac{\sum_{i=j_2+1}^{l} m_i}{\sum_{i=j_2+1}^{l} T_i} \tag{4.25}
\]

4.4 Compound Poisson Information Criterion

In a compound Poisson model, the rate at which the event occurs is of interest in application [14]. For the case of at most a single change point, Plummer proposes a new modified
information criterion for compound Poisson data, known as Compound Poisson Information Criterion. Under CPIC, the penalty term, denoted by $\kappa_p$, where $p$ is the number of change points, on each location is reduced for positions which create changes in $\lambda$. The penalty term for the single change point model is given by:

$$\kappa_1 = \left[ 2 \text{dim}(\theta) + \left( 1 + \sum_{i=1}^{2} |\hat{\lambda} - \hat{\lambda}_i| \right)^{-1} \right] \log(n) \text{, for } n \text{ observations} \quad (4.26)$$

The general form of the CPIC equation for the alternative hypothesis of $p$ change points is as follows:

$$CPIC(j_1, j_2, \ldots, j_p) = -2 \log L(\hat{\theta}) + \kappa_p \quad , \quad p \in \{1, \ldots, l - 1\} \quad (4.27)$$

Where $\hat{\theta}$ is the MLEs of $\theta$.

The equations for $CPIC(n)$, $CPIC(j)$, and $CPIC(j_1, j_2)$ can be applied to test the null and alternative hypotheses.
Under $H_0$ as stated in section (3.1), Plummer gives $CPIC(n)$ as [14]:

$$CPIC(n) = -2\log L_0(\hat{\theta}) + \kappa_0$$

$$= -2\log L_0(\hat{\mu}, \hat{\lambda}) + 2\log(n)$$

$$= M_A log(2\pi \sigma^2) + \sum_{i \in A} \log(m_i) + \frac{1}{\sigma^2} \sum_{i \in A} \frac{(x_{t_i} - m_i \hat{\mu})^2}{m_i}$$

$$- 2 \sum_{i=1}^{l} m_i \log(T_i) + 2 \sum_{i=1}^{l} \log(m_i!) - 2\log(\hat{\lambda}) \sum_{i=1}^{l} m_i$$

$$+ 2\hat{\lambda} \sum_{i=1}^{l} T_i + 2\log(n) \quad (4.28)$$

For a single change point, $j$, $j = 1, 2, \ldots, l - 1$, Plummer gives $H_1$ as [14]:

$$CPIC(j) = -2\log L_1(\hat{\theta}) + \kappa_1$$

$$= -2\log L_1(\hat{\theta}) + \left[ 2\dim(\theta) + \left( 1 + \sum_{i=1}^{2} |\hat{\lambda} - \hat{\lambda}_i|^{-1} \right) \right] \log(n)$$

$$= M_A log(2\pi \sigma^2) + \sum_{i \in A} \log(m_i) + \frac{1}{\sigma^2} \sum_{i \in A_j} \frac{(x_{t_i} - m_i \hat{\mu}_1)^2}{m_i}$$

$$- 2 \sum_{i=1}^{l} m_i \log(T_i) + 2 \sum_{i=1}^{l} \log(m_i!) - 2\log(\hat{\lambda}_1) \sum_{i=1}^{j} m_i$$

$$+ 2\hat{\lambda}_1 \sum_{i=1}^{j} T_i + \frac{1}{\sigma^2} \sum_{i \in A_j} \frac{(x_{t_i} - m_i \hat{\mu}_2)^2}{m_i} - 2\log(\hat{\lambda}_2) \sum_{i=j+1}^{l} m_i$$

$$+ 2\hat{\lambda}_2 \sum_{i=j+1}^{l} T_i + \left[ 2\dim(\theta) + \left( 1 + \sum_{i=1}^{2} |\hat{\lambda} - \hat{\lambda}_i|^{-1} \right) \right] \log(n) \quad (4.29)$$
Under the alternative hypothesis for two change points, \( CPIC(j_1, j_2) \) is derived for 1 ≤ \( j_1 < j_2 \) ≤ \( l - 1 \).

\[
CPIC(j_1, j_2) = -2\log L_1(\hat{\theta}) + \kappa_2
\]
\[
= -2\log L_1(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3) + \left[3\dim(\theta) + \left(1 + \sum_{i=1}^{3} |\hat{\lambda} - \hat{\lambda}_i|\right)^{-1}\right] \log(n)
\]
\[
= MA \log(2\pi \sigma^2) + \frac{1}{\sigma^2} \left[\sum_{i \in A_{j_1}} (x_{t_i} - \mu_1)^2 + \sum_{i \in A_{j_2}} (x_{t_i} - \mu_2)^2 + \sum_{i \in A_l} (x_{t_i} - \mu_3)^2\right]
\]
\[
-2 \sum_{i=1}^{j_1} \log(T_i) - 2 \sum_{i=j_1+1}^{j_2} \log(\lambda_1) - 2 \sum_{i=j_1+1}^{j_2} \log(\lambda_2) - 2 \sum_{i=j_2+1}^{l} \log(\lambda_3)
\]
\[
+ 2\lambda_1 \sum_{i=1}^{j_1} T_i + 2\lambda_2 \sum_{i=j_1+1}^{j_2} T_i + 2\lambda_3 \sum_{i=j_2+1}^{l} T_i
\]
\[
+ \sum_{i \in A} \log(m_i) + 2\log \sum_{i=1}^{l} \log(m_i!)
\]
\[
+ \left[3\dim(\theta) + \left(1 + \sum_{i=1}^{3} |\hat{\lambda} - \hat{\lambda}_i|\right)^{-1}\right] \log(n)
\]

(4.30)

Using \( CPIC(n) \), \( CPIC(j) \), and \( CPIC(j_1, j_2) \) we can test the null hypothesis.

To test the hypothesis of no change versus at most two changes, first define

\[
CPIC(j^*) = \min_{1 \leq j \leq l-1} \{CPIC(j)\}
\]

(4.31)

\[
CPIC(j_1^*, j_2^*) = \min_{1 \leq j_1 < j_2 \leq l-1} \{CPIC(j_1, j_2)\}
\]

(4.32)
To test $H_0$, reject if

$$CPIC(n) > \min\{CPIC(j^*), CPIC(j_1^*, j_2^*)\}. \quad (4.33)$$

Fail to reject $H_0$ if

$$CPIC(n) \leq \min\{CPIC(j^*), CPIC(j_1^*, j_2^*)\}. \quad (4.34)$$

If $H_0$ is rejected, the estimated location(s) of the change point(s) is given by:

$$CPIC(\hat{j}) = \min_{1 \leq j \leq l-1} CPIC(j) \quad (4.35)$$

if $CPIC(j^*) \leq CPIC(j_1^*, j_2^*)$, and by;

$$CPIC(\hat{j}_1, \hat{j}_2) = \min_{1 \leq j_1 < j_2 \leq l-1} CPIC(j_1, j_2) \quad (4.36)$$

if $CPIC(j^*) > CPIC(j_1^*, j_2^*)$. 

Chapter 5

Simulation Study and Application to aCGH Data

5.1 Overview

To assess and confirm the accuracy of the CPIC method developed to identify at most two change points, a thorough simulation study was performed using R statistical software. To simulate a compound Poisson process a set of normal observations were randomly generated and paired with a set of randomly generated exponential variables. This simulates the case of one observation per interval and the exponential variables represent the distances between observations. The simulation randomly generated data with at most two changes in parameter.

For simulations with a change point or points, three levels of change were considered: small change, medium change, and large change; as well as three positions for the change(s) towards the front, middle, and end of the data sets. For the case of no change, varying levels of parameter values were considered. All combinations of number of change, parameter level, and position were repeated 1000 times.
Once the accuracy of the method was confirmed using simulation, the algorithm was run using aCGH (Array Comparative Genomic Hybridization) cell line data. This data set contains two random variables; a normal variable and a Poisson variable. The data also contained a non random variable which are the lengths of the intervals. Since some intervals contained more than one observation, the R script had to be modified to evaluate these data sets.

5.2 Simulation Study

5.2.1 Simulations with Two Change Points

The first part of the simulation involved a variety of tests in which the parameters changed two times, thus, there existed two change points. In these simulations, there were two sets of parameters. Initially, the mean and rate were set to \( \mu_0 = 0, \lambda_0 = 1 \). After the location of the first change point, \( j_1 \), the parameters would change to \( (\mu_1, \lambda_1) \). The data would follow \( (\mu_1, \lambda_1) \) up to the location of the second change point, \( j_2 \). At \( j_2 + 1 \), the parameters would change back to \( (\mu_0, \lambda_0) = (0, 1) \).

The values for \( (\mu_1, \lambda_1) \) were chosen to be \( (2, 2), (3, 3), \) and \( (3.5, 4) \). Simulations were performed on sample sizes of \( N = 40, N = 60, N = 100, \) and \( N = 200 \). The locations of \( j_1, j_2 \) were set to the pairs of \( n/4, 3n/8; n/2, 5n/8; 3n/4, 7n/8 \). These pairs were chosen to show that the algorithm is consistent for change points located at the front, middle, and end of the data sets. All combinations of parameters, sample size, and change point locations were run in the simulations. In addition to the locations previously listed, for \( N = 60 \) and \( N = 100 \), locations of \( j_1, j_2 \) were also set at \( j_1 = n/4, j_2 = n/4 + 5; j_1 = n/2, j_2 = n/2 + 5; j_1 = 3n/4, j_2 = 3n/4 + 5 \). These locations were simulated for each set of \( (\mu_1, \lambda_1) \). A total of 54 sets simulations were run with two change points. Figure 5.1 is a plot of the compound
Poisson random variables for one of the 1000 repetitions for $N=200$, $\mu_1 = 3.5$, $\lambda_1 = 4$, $j_1 = 150$, $j_2 = 175$. The circled points are the change points.

![Simulated Compound Poisson Data for N=200](image)

Figure 5.1: Simulated Compound Poisson Data for $N=200$

In Figure 5.1, the change in the parameters at $j_1 = 150$ can be seen. At $j_2 + 1 = 176$ the parameters change to $\mu_0 = 0$, $\lambda_0 = 1$. The algorithm correctly identified $j_1$ and $j_2$ in this plot. As Table 5.4 indicates, for $N=200$, $\mu_1 = 3.5$, $\lambda_1 = 4$, the correct model was chosen at a rate of 1.0.

The results, mean position selected, and MSE for the simulations organized by sample size and location of the change points are given in Tables 5.1, 5.2, and 5.3.
Table 5.1: Results Table for \((\mu_1, \lambda_1) = (2, 2)\)

<table>
<thead>
<tr>
<th>N</th>
<th>(f)</th>
<th>(n/4)</th>
<th>(n/2)</th>
<th>(3n/4)</th>
<th>(3n/8)</th>
<th>(5n/8)</th>
<th>(7n/8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>.54</td>
<td>.507</td>
<td>.528</td>
<td>.556</td>
<td>.528</td>
<td>.503</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>10.894</td>
<td>19.19</td>
<td>27.075</td>
<td>16.42</td>
<td>25.069</td>
<td>33.087</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>22.782</td>
<td>20.094</td>
<td>62.189</td>
<td>29.748</td>
<td>16.871</td>
<td>35.385</td>
</tr>
<tr>
<td>60</td>
<td>.574</td>
<td>.607</td>
<td>.582</td>
<td>.609</td>
<td>.615</td>
<td>.562</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>15.388</td>
<td>29.641</td>
<td>43.493</td>
<td>23.855</td>
<td>37.974</td>
<td>52.2</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>18.698</td>
<td>8.357</td>
<td>49.323</td>
<td>23.529</td>
<td>10.35</td>
<td>19.526</td>
</tr>
<tr>
<td>100</td>
<td>.645</td>
<td>.645</td>
<td>.629</td>
<td>.628</td>
<td>.639</td>
<td>.604</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>25.191</td>
<td>49.996</td>
<td>74.589</td>
<td>38.178</td>
<td>63.051</td>
<td>87.784</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>8.63</td>
<td>2.322</td>
<td>16.659</td>
<td>6.228</td>
<td>2.629</td>
<td>6.586</td>
</tr>
<tr>
<td>200</td>
<td>.687</td>
<td>.678</td>
<td>.652</td>
<td>.664</td>
<td>.655</td>
<td>.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>49.941</td>
<td>99.972</td>
<td>150.014</td>
<td>75.082</td>
<td>125.07</td>
<td>174.977</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>1.669</td>
<td>1.344</td>
<td>1.862</td>
<td>2.634</td>
<td>1.518</td>
<td>2.517</td>
</tr>
</tbody>
</table>
Table 5.2: Results Table for \((\mu_1, \lambda_1) = (3, 3)\)

<table>
<thead>
<tr>
<th>N</th>
<th>(f)</th>
<th>(n/4)</th>
<th>(n/2)</th>
<th>(3n/4)</th>
<th>(3n/8)</th>
<th>(5n/8)</th>
<th>(7n/8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>.834</td>
<td>.851</td>
<td>.863</td>
<td>.867</td>
<td>.867</td>
<td>.787</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>10.01</td>
<td>19.9</td>
<td>29.728</td>
<td>15.067</td>
<td>25.024</td>
<td>34.91</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>1.324</td>
<td>1.172</td>
<td>5.424</td>
<td>1.285</td>
<td>.95</td>
<td>2.03</td>
</tr>
<tr>
<td>60</td>
<td>.862</td>
<td>.868</td>
<td>.874</td>
<td>.866</td>
<td>.857</td>
<td>.837</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>14.988</td>
<td>29.977</td>
<td>44.927</td>
<td>22.977</td>
<td>38.025</td>
<td>53.002</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>.444</td>
<td>.273</td>
<td>1.767</td>
<td>.277</td>
<td>.259</td>
<td>.426</td>
</tr>
<tr>
<td>100</td>
<td>.864</td>
<td>.859</td>
<td>.881</td>
<td>.87</td>
<td>.872</td>
<td>.864</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>24.99</td>
<td>49.986</td>
<td>74.977</td>
<td>38.018</td>
<td>62.991</td>
<td>87.994</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>.224</td>
<td>.3</td>
<td>.191</td>
<td>.229</td>
<td>.225</td>
<td>.248</td>
</tr>
<tr>
<td>200</td>
<td>.849</td>
<td>.886</td>
<td>.885</td>
<td>.891</td>
<td>.876</td>
<td>.871</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>49.996</td>
<td>100.004</td>
<td>149.968</td>
<td>75.012</td>
<td>124.995</td>
<td>174.993</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>.242</td>
<td>.19</td>
<td>.204</td>
<td>.23</td>
<td>.197</td>
<td>.267</td>
</tr>
</tbody>
</table>
Table 5.3: Results Table for \((\mu_1, \lambda_1) = (3.5, 4)\)

<table>
<thead>
<tr>
<th></th>
<th>Change Position One</th>
<th>Change Position Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>(n/4)</td>
<td>(n/2)</td>
</tr>
<tr>
<td>40</td>
<td>.909</td>
<td>.929</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>9.994</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>.15</td>
</tr>
<tr>
<td>60</td>
<td>.938</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>14.976</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>.126</td>
</tr>
<tr>
<td>100</td>
<td>.923</td>
<td>.925</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>24.984</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>.12</td>
</tr>
<tr>
<td>200</td>
<td>.944</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>49.994</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>.068</td>
</tr>
</tbody>
</table>

In Tables 5.1, 5.2, and 5.3 a trend towards higher accuracy as N increases and as the difference in the parameters increases can be seen. In table 5.3, for N=200 the frequency the correct location of the change points were identified were all above .9 and the MSE all below 1.0. Also, the mean location calculated were all within .01 of the actual location of the change points. In contrast, in Table 5.1, at the lowest level of change in parameter, \(\mu_1 = 2, \lambda_1 = 2\), and N=40, the algorithm is less accurate in detection. From observation, the algorithm will be more accurate in detecting change points in larger sample sizes with greater changes in the parameters.
Table 5.4: Frequency of Selecting Two Change Point Model and Positions

<table>
<thead>
<tr>
<th>$(\mu_1, \lambda_1)$</th>
<th>$N$</th>
<th>Positions</th>
<th>$f_{\text{model} \ 2}$</th>
<th>$f_{\text{j1, j2}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(2, 2)$</td>
<td>40</td>
<td>$(n/4, 3n/8)$</td>
<td>.883</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>.881</td>
<td>.314</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>.872</td>
<td>.331</td>
</tr>
<tr>
<td>$(3, 3)$</td>
<td></td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>.998</td>
<td>.698</td>
</tr>
<tr>
<td>$(3.5, 4)$</td>
<td></td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.833</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.852</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.827</td>
</tr>
<tr>
<td>$(2, 2)$</td>
<td>60</td>
<td>$(n/4, 3n/8)$</td>
<td>.977</td>
<td>.361</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>.976</td>
<td>.397</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>.974</td>
<td>.361</td>
</tr>
<tr>
<td>$(3, 3)$</td>
<td></td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.747</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.749</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.74</td>
</tr>
<tr>
<td>$(3.5, 4)$</td>
<td></td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.861</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.854</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.866</td>
</tr>
<tr>
<td>$(2, 2)$</td>
<td>100</td>
<td>$(n/4, 3n/8)$</td>
<td>.998</td>
<td>.409</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.447</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.407</td>
</tr>
<tr>
<td>$(3, 3)$</td>
<td></td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.749</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.752</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.765</td>
</tr>
<tr>
<td>$(3.5, 4)$</td>
<td></td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.849</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.868</td>
</tr>
<tr>
<td>$(2, 2)$</td>
<td>200</td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.446</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.444</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.411</td>
</tr>
<tr>
<td>$(3, 3)$</td>
<td></td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.782</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.771</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.779</td>
</tr>
<tr>
<td>$(3.5, 4)$</td>
<td></td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.877</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.857</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.88</td>
</tr>
</tbody>
</table>
Table 5.5: Accuracy of Model Selection for Second Change Point at Location $j_1+5$

<table>
<thead>
<tr>
<th>$(\mu_1, \lambda_1)$</th>
<th>$N$</th>
<th>Positions</th>
<th>$f_{model\ 2}$</th>
<th>$f_{j_1, j_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(2, 2)$</td>
<td>40</td>
<td>$(n/4, 3n/8)$</td>
<td>.883</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>.881</td>
<td>.314</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>.872</td>
<td>.331</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>$(n/4, n/4 + 5)$</td>
<td>.871</td>
<td>.335</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, n/2 + 5)$</td>
<td>.867</td>
<td>.349</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 3n/4 + 5)$</td>
<td>.82</td>
<td>.325</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>$(n/4, n/4 + 5)$</td>
<td>.816</td>
<td>.311</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, n/2 + 5)$</td>
<td>.81</td>
<td>.346</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 3n/4 + 5)$</td>
<td>.8</td>
<td>.348</td>
</tr>
<tr>
<td>$(3, 3)$</td>
<td>40</td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>.998</td>
<td>.698</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>$(n/4, n/4 + 5)$</td>
<td>1.0</td>
<td>.753</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, n/2 + 5)$</td>
<td>.998</td>
<td>.712</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 3n/4 + 5)$</td>
<td>.997</td>
<td>.729</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>$(n/4, n/4 + 5)$</td>
<td>1.0</td>
<td>.747</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, n/2 + 5)$</td>
<td>.999</td>
<td>.739</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 3n/4 + 5)$</td>
<td>.999</td>
<td>.728</td>
</tr>
<tr>
<td>$(3.5, 4)$</td>
<td>40</td>
<td>$(n/4, 3n/8)$</td>
<td>1.0</td>
<td>.833</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, 5n/8)$</td>
<td>1.0</td>
<td>.852</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 7n/8)$</td>
<td>1.0</td>
<td>.827</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>$(n/4, n/4 + 5)$</td>
<td>1.0</td>
<td>.847</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, n/2 + 5)$</td>
<td>1.0</td>
<td>.851</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 3n/4 + 5)$</td>
<td>1.0</td>
<td>.854</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>$(n/4, n/4 + 5)$</td>
<td>1.0</td>
<td>.848</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(n/2, n/2 + 5)$</td>
<td>1.0</td>
<td>.865</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(3n/4, 3n/4 + 5)$</td>
<td>1.0</td>
<td>.862</td>
</tr>
</tbody>
</table>
Table 5.4 displays the frequency the algorithm selects the two change point model and the frequency that the locations of both change points are detected correctly. For the largest sample size, N=200, the algorithm chose the correct model at a rate of 1.0. Also, for the greatest change in the parameters, $\mu_1 = 3.5$, $\lambda_1 = 4$, the algorithm correctly detected the location of both change points at a rate of at least .833. For the smallest sample size and least change in the parameters, the rate of correct model selection was at least .872. For this sample size and parameters, the rate of correct detection of both change points location was not greater than .35. Table 5.4 leads to the conclusion that larger sample size and greater change in the parameters will result in more accurate model selection and correct detection of both change points.

For sample sizes N=40, 60, and 100 simulations where the position of the second change point was five greater than the first change point were run. These results are summarized in Table 5.5. Comparing Tables 5.4 and 5.5 for N=60 and N= 100 it is noticeable that the rate that the algorithm correctly identifies the location of both change points tends to be lower in the case where $j_2$ is located at $j_1 + 5$. This is intuitive from a CPIC approach since an increase in the sample size is providing more information about the initial parameters but no additional about the alternative parameters.

### 5.2.2 Simulations with Zero Change Points

Simulations in which there was no change in the parameters were also performed. The objective was to assess the decision making ability of the algorithm when the data contained zero change points. Twelve sets of simulations without any change in parameter were run. The sample sizes chosen were $N = 40$, $N = 60$, $N = 100$, and $N = 200$. Three sets of parameters, $(\mu_0, \lambda_0)$, for each sample size were chosen; $(0, 1)$, $(0, 3)$, and $(3.5, 4)$. The results are given in Table 5.6. In Table 5.6, the frequency that the algorithm chose the correct
model had a minimum of .913 and a maximum of .988. The minimum frequency of correct model selection was in the smallest sample size, N=40. The maximum frequency of correct model selection was in the largest sample size, N=200. From these simulation results, we can conclude that as the sample size increases, the frequency of correct model selection will also increase.

Table 5.6: Null Simulation Results

<table>
<thead>
<tr>
<th>N</th>
<th>µ</th>
<th>λ</th>
<th>f_{null}</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0</td>
<td>1</td>
<td>.916</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>.913</td>
</tr>
<tr>
<td></td>
<td>3.5</td>
<td>4</td>
<td>.916</td>
</tr>
<tr>
<td>60</td>
<td>0</td>
<td>1</td>
<td>.951</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>.941</td>
</tr>
<tr>
<td></td>
<td>3.5</td>
<td>4</td>
<td>.934</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>1</td>
<td>.969</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>.967</td>
</tr>
<tr>
<td></td>
<td>3.5</td>
<td>4</td>
<td>.964</td>
</tr>
<tr>
<td>200</td>
<td>0</td>
<td>1</td>
<td>.977</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>.982</td>
</tr>
<tr>
<td></td>
<td>3.5</td>
<td>4</td>
<td>.988</td>
</tr>
</tbody>
</table>

A false positive result in simulation is shown for the exponential random variables in Figure 5.2. Upon observation of the scatterplot it can be seen that an anomoly in the randomly generated exponential variable is the cause of the false positive. For this simulation N=40, \( \mu_0 = 0, \lambda_0 = 1 \). The algorithm selected the single change point model in this instance. Figure 5.3 shows the normal random variables from the same simulation as in Figure 5.2. However, the plot clearly shows that no change point has occured.
5.2.3 Simulations with a Single Change Point

Plummer extensively studied the single change point problem using the CPIC approach [14]. The algorithm developed in this paper was also used for the case of a single change point. For the case of a single change point, the algorithm chose the correct model at a rate in about 60% of the simulations, and selected the two change point model in the remaining simulations. Upon further investigation it appears that the model misspecifications were due to anomalies in the sequence of the exponentials being generated.
5.3 Application to aCGH Data Set

The next step in the process was to apply the information criterion to sets of gene expression data. The gene expression data is real world data collected by medical professionals for the purpose of cancer research. These data sets contain the log₂ ratio of the gene expressions, which are known to be approximately normal, the distances between gene locations and the number of genes measured on each interval. To maintain as much information as possible the intervals containing more than one gene were held to a minimum. Since the location of the gene markers were selected for their biological significance and not through a random process, the lengths of the intervals varied from a range of close to zero to over 23000. A notable example of such an extreme change in length was an interval of length 0.0005 adjacent to an interval of length 379.952. However, the log₂ ratio for these two positions show no indication of a change in mean. From a statistician’s perspective, it was clear that the approach taken in this paper may cause the algorithm to detect a change at such locations. Another difference in the aCGH data is that within the cell line there are physical barriers [14]. Observations were taken at these barriers, which means that these observations were in an interval with a length of zero. From the CPIC equation, this would cause a domain error. Therefore, data preparation was necessary to prevent domain errors. A minimalistic approach toward data preparation was followed to preserve the data as close to the manner in which the observations were sampled as possible. Intervals with length less than one were combined with the adjacent interval.
### Table 5.7: aCGH Application Results Table

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Chrom.</th>
<th>C.P.s Found</th>
<th>False Pos. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM01524</td>
<td>6</td>
<td>2</td>
<td>.174</td>
</tr>
<tr>
<td>GM03134</td>
<td>8</td>
<td>2</td>
<td>.348</td>
</tr>
<tr>
<td>GM05296</td>
<td>10</td>
<td>2</td>
<td>.391</td>
</tr>
<tr>
<td>GM05296</td>
<td>11</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>GM13031</td>
<td>17</td>
<td>2</td>
<td>.435</td>
</tr>
<tr>
<td>GM01535</td>
<td>5</td>
<td>1</td>
<td>.304</td>
</tr>
<tr>
<td>GM01750</td>
<td>9</td>
<td>1</td>
<td>.348</td>
</tr>
<tr>
<td>GM01750</td>
<td>14</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GM03563</td>
<td>3</td>
<td>1</td>
<td>.217</td>
</tr>
<tr>
<td>GM03563</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GM04435</td>
<td>5</td>
<td>1</td>
<td>.348</td>
</tr>
<tr>
<td>GM04435</td>
<td>16</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>GM07081</td>
<td>7</td>
<td>1</td>
<td>1*</td>
</tr>
<tr>
<td>GM13330</td>
<td>1</td>
<td>1</td>
<td>.391</td>
</tr>
<tr>
<td>GM13330</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

*Denotes incorrect model selection

![Chromosome 6](image.png)

Figure 5.4: Chromosome with Two Change Points Correctly Identified
Table 5.7 summarizes the results of the application to the gene expression data. Three cell lines contained one chromosome with two change points. In two of these cell lines, the change points were identified correctly. In the other two these cell lines, the change points were not found. One cell line had two chromosomes each with two change points. In each of these chromosomes, the two change point model was selected, however, only one of the two change point positions were correct in each chromosome. Note that the table shows when the location of the change point was identified at the exact location. Six cell lines contained one or more chromosome with a single change point. Of these, there were ten chromosomes with a single change point. Five times the change point was located. In one of the cell lines, the single change point was located, but the algorithm falsely detected a second change point; this is denoted by an asterisk in Table 5.7 and illustrated in Figure 5.5.

![Figure 5.5: Single Change Point Detected, Two Change Point Model Selected](image1)

![Figure 5.6: False Positives From Outlier](image2)
The false positive rates had a minimum of .174 and a maximum of .435. As shown in Figure 5.6, outliers can cause a false positive detection. In Figure 5.7 it can be shown that outliers do not always cause false positives. The algorithm detected the single change point even in the presence of an outlier. From observation of the algorithm outputs, it seems as though there were cases where the algorithm detected a change point in close proximity to the actual change point. In such a case, this is still tabulated as a false positive. As previously noted, the data samples were taken at locations of biological significance, thus, the interval lengths varied greatly. From the CPIC approach, interval lengths that vary greatly where no change in expression level occurs, could be the reason for many of the false positives.

In conclusion, the simulation study and application to the gene expression data was useful in the study of the CPIC approach. It is clear that the CPIC performed better in simulation than in the application to the aCGH data. This was due to the sensitivity of the CPIC in changes to the rate of occurrence as well as changes in the normal variables. The randomly generated data sets better conformed to the method used in change point location detection, rather than the aCGH data sets. From these results, the CPIC method was found to be useful in detecting changes in mean, rate of occurrence, or both.
Appendix A

R Code

A.1 Simulation R Code

# Written by Jake Roach and Paul Plummer #

# List of Variables #
# N is the number of observations in each sample #
# B1 is the location of the first change #
# B2 is the location of the second change #
# Res is the matrix of the results of the 1000 samples #
# Y is the matrix with the 1000 samples #
# k is the current position being searched as a change point in the loop #
# S0 is the summation of the squared deviations of the normal variables from ybar #
# S1 is the summation of the first k squared deviations of the normal variables #
# ybar0 is the average of the normal observations under the null hypothesis#
# ybar1 is the average of the first k normal variables#
# ybar2 is the average of the last N-k normal variables#
# ybar3 is the average of the N-(k2+1) normal variables#
# M1 is equal to k #
# M2 is equal to N-k #
# M3 is equal to N-k2 #
# i is a loop parameter #
# j is the loop parameter for the data pair #
# V is the base variance for the normal model under the null hypothesis #
# t0 is the total of the exponential lengths of the intervals #
# t1 is the total of the interval lengths from 1:k #
# t2 is the total of the interval lengths from k+1:N #
# t3 is the total of the interval lengths from k2+1:N #
# P0 is the CPIC for \(-2\log(L0)\); L0 is the likelihood function under H0 #
# P1 is the CPIC for \(-2\log(L1)\); L1 is the likelihood function under H1 #
# M1/t1 is lambda hat 1 #
# M2/t2 is lambda hat 2 #
# M3/t3 is lambda hat 3 #

library(MASS)
N=100 ;
B1=75;
B2=88;
mu1=0;
mu2=2;
mu3=0;
V=1;
lambda1=1;
lambda2=2;
lambda3=1;
#Simulation#
Y<-matrix(0,N,2000);
for (i in 1:1000) {
    Y[1:B1,2*i]<-rexp(B1,lambda1);
    Y[(B1+1):B2,2*i]<-rexp(B2-B1,lambda2);
    Y[1:B1,2*i-1]<-rnorm(B1,mu1,sqrt(V));
    Y[(B1+1):B2,2*i-1]<-rnorm(B2-B1,mu2,sqrt(V));
    Y[(B2+1):N,2*i]<-rexp(N-B2,lambda3);
}

#Start Search#
Res <-matrix(0,1001,19);
for(j in 1:1000) {

    #Calculation of the null hypothesis#
ybar0<-sum(Y[1:N,2*j-1])/N;
t_0<-sum(Y[1:N,2*j]);
S0=0;
n0=0;
for (i in 1:N) S0=S0+(Y[i,2*j-1]-ybar0)^2; n0=n0+log(Y[i,2*j])
P0 <-N*log(2*pi*V)+S0/V-2*N*log(N/t_0)-2*n0+2*N+2*log(N);
Res[j+1,2]=P0; Res[j+1,3]=ybar0; Res[j+1,4]=N/t_0;

    #Search for a single change point#    for(k in 1:(N-1)) {
        S1 = 0;
ybar1<-sum(Y[1:k,2*j-1])/k;
ybar2<-sum(Y[(k+1):N,2*j-1])/(N-k);
APPENDIX A. R CODE

M1<-k;
M2<-N-k;
t_1 <-sum(Y[1:k,2*j]);
t_2<-sum(Y[(k+1):N,2*j]);
for(i in 1:k) S1 = S1 +(Y[i,2*j-1]-ybar1)^ 2;
for(i in (k+1):N) S1 = S1 +(Y[i,2*j-1]-ybar2)^ 2;
P1<-N*log(2*pi*V)+S1/V-2*M1*log(M1/t_1)-2*M2*log(M2/t_2)
-2*n0+2*N+(4+1/(1+abs(N/t_0-M1/t_1)+abs(N/t_0-M2/t_2)))*log(N);
if(k==1) {
  Res[j+1,5]=k; Res[j+1,6]=P1; Res[j+1,7]=ybar1;
  Res[j+1,8]=ybar2; Res[j+1,9]=M1/t_1; Res[j+1,10]=M2/t_2;
}
if(P1 < Res[j+1,6]) {
  Res[j+1,5]=k; Res[j+1,6]=P1; Res[j+1,7]=ybar1;
  Res[j+1,8]=ybar2; Res[j+1,9]=M1/t_1; Res[j+1,10]=M2/t_2;
}

#Search for two change points#
for(k in 1:(N-2)) {
  for(k2 in (k+1):(N-1)) {
    S1=0;
ybar1<-sum(Y[1:k,2*j-1])/k;
ybar2<-sum(Y[(k+1):k2,2*j-1])/(k2-k);
ybar3<-sum(Y[(k2+1):N,2*j-1])/(N-k2);
  }
}
M1 <- k;
M2 <- k2 - k
M3 <- N - k2

t.1 <- sum(Y[1:k, 2*j]);
t.2 <- sum(Y[(k+1):k2, 2*j]);
t.3 <- sum(Y[(k2+1):N, 2*j]);

for(i in 1:k) S1 = S1 + (Y[i, 2*j-1] - ybar1)^2;
for(i in (k+1):k2) S1 = S1 + (Y[i, 2*j-1] - ybar2)^2;
for(i in (k2+1):N) S1 = S1 + (Y[i, 2*j-1] - ybar3)^2;

P2 <- N*log(2*pi*V) + S1/V - 2*M1*log(M1/t1) - 2*M2*log(M2/t2) - 2*M3*log(M3/t3)

-2*n0 + 2*N + (6 + 1/(1 + abs(N/t0 - M1/t1) + abs(N/t0 - M2/t2) + abs(N/t0 - M3/t3))) * log(N);
if(k2 == 2) {
  Res[j+1,11] = k; Res[j+1,12] = k2; Res[j+1,13] = P2;
  Res[j+1,14] = ybar1; Res[j+1,15] = ybar2; Res[j+1,16] = ybar3;
}
if(P2 < Res[j+1,13]) {
  Res[j+1,11] = k; Res[j+1,12] = k2; Res[j+1,13] = P2;
  Res[j+1,14] = ybar1; Res[j+1,15] = ybar2; Res[j+1,16] = ybar3;
}
CPIC[j] <- min(Res[j+1,2], Res[j+1,6], Res[j+1,13]);
if(CPIC[j] == Res[j+1,6]) Res[j+1,1] <- -1;
if(CPIC[j] == Res[j+1,13]) Res[j+1,1] <- -2;};
APPENDIX A. R CODE

Headings<-t(c("Model_Selected","CPIC(m)","Mu0","lambda0",
"j","CPIC(j)","Mu1","Mu2","lambda1","lambda2",
"j.1","j.2","CPIC(j.1,j.2)","Mu1","Mu2","Mu3",
"lambda1","lambda2","lambda3")); Res[1,]<-Headings;

#use write.matrix to create an information file and an output file#
library(MASS)

D <- read.table(file=file.choose(), header=TRUE)
attach(D)

GL <- dim(D)[1];
Mark <- array(D[,1]);
cs = 1;
ce = 0;

Y <- matrix(0, GL, 3);
Y[,1] <- D[,3];
Y[,2] <- D[,4];
Y[,3] <- D[,5];

Res <- matrix(0, 24, 18);

V <- var(Y[,1]);
Y[,1] = Y[,1]/sqrt(V);
c = 1;

for (C in 1:23) {
  if (C != 23) {
    while (D[ce+1,2] == C) ce = ce + 1;
  }
  else ce = GL; NS = sum(Y[cs:ce,3]);

  # Start Search#
  ybar0 = sum(Y[cs:ce,1])/N;
  t.0 = sum(Y[cs:ce,2]);
  S0 = 0;
  n0 = 0;
  n1 = 0;
APPENDIX A. R CODE

\begin{verbatim}
f=0;
for (i in cs:ce) {S0=S0+((Y[i,1]-Y[i,3]*ybar0) ^ 2)/Y[i,3] ;
n0=n0+Y[i,3]*log(Y[i,2]); n1=n1+log(Y[i,3]); f=f+log(factorial(Y[i,3]))};
P0<- (ce-cs+1)*log(2*pi)+n1+S0-2*n0-2*N*log(N/t0)+2*N+2*f+2*log(ce-cs+1);

for(k in cs:(ce-1)) { S1 = 0;
M1<-sum(Y[cs:k,3]);
M2<-sum(Y[(k+1):ce,3]);
ybar1<-sum(Y[cs:k,1])/M1;
ybar2<-sum(Y[(k+1):ce,1])/M2;
t.1<-sum(Y[cs:k,2]);
t.2<-sum(Y[(k+1):ce,2]);
for(i in cs:k) S1 = S1 +((Y[i,1]-Y[i,3]*ybar1)ˆ2)/Y[i,3];
for(i in (k+1):ce) S1 = S1 +((Y[i,1]-Y[i,3]*ybar2)ˆ2)/Y[i,3];
P1<-((ce-cs+1)*log(2*pi)+n1+S1-2*n0-2*M1*log(M1/t1)-2*M2*log(M2/t2)+2*N+2*f
+(4+c/(1+abs(N/t0-M1/t1)+abs(N/t0-M2/t2))))*log(ce-cs+1);

if(k==cs) {Res[C+1,2]=P0; Res[C+1,3]=D[k,1]; Res[C+1,4]=P1; Res[C+1,5]=ybar1;
Res[C+1,6]=M1/t1; Res[C+1,7]=ybar2; Res[C+1,8]=M2/t2};
if(P1 < Res[C+1,4]) {Res[C+1,3]=D[k,1]; Res[C+1,4]=P1; Res[C+1,5]=ybar1;
Res[C+1,6]=M1/t1; Res[C+1,7]=ybar2; Res[C+1,8]=M2/t2}; }

for(k in cs:(ce-2)) { for(k2 in (k+1):(ce-1)){
S1 = 0;
M1<-sum(Y[cs:k,3]);
M2<-sum(Y[(k+1):k2,3]);
\end{verbatim}
APPENDIX A. R CODE

\begin{verbatim}
M3 <- sum(Y[(k2+1):ce,3]);
ybar1 <- sum(Y[cs:k,1])/M1;
ybar2 <- sum(Y[(k+1):k2,1])/M2;
ybar3 <- sum(Y[(k2+1):ce,1])/M3;
t.1 <- sum(Y[cs:k,2]);
t.2 <- sum(Y[(k+1):k2,2]);
t.3 <- sum(Y[(k2+1):ce,2]);

for(i in cs:k) S1 = S1 + (Y[i,1] - Y[i,3]*ybar1)^2)/Y[i,3];
for(i in (k+1):k2) S1 = S1 + (Y[i,1] - Y[i,3]*ybar2)^2)/Y[i,3];
for(i in (k2+1):ce) S1 = S1 + (Y[i,1] - Y[i,3]*ybar3)^2)/Y[i,3];

P2 <- -(ce-cs+1)*log(2*pi)+n1+S1-2*n0-2*M1*log(M1/t.1)
       -2*M2*log(M2/t.2)-2*M3*log(M3/t.3)+2*N+2*f
       +(6+1/(1+abs(N/t.0-M1/t.1)+abs(N/t.0-M2/t.2)+abs(N/t.0-M2/t.3)))
       *log(ce-cs+1);

if(k==cs) {Res[C+1,9]=D[k,1]; Res[C+1,10]=D[k2,1];
             Res[C+1,11]=P2; Res[C+1,12]=ybar1; Res[C+1,13]=M1/t.1;
             Res[C+1,14]=ybar2; Res[C+1,15]=M2/t.2;
             Res[C+1,16]=ybar3; Res[C+1,17]=M3/t.3};
if(P2 < Res[C+1,11])
{Res[C+1,9]=D[k,1]; Res[C+1,10]=D[k2,1]; Res[C+1,11]=P2; Res[C+1,12]=ybar1;
             Res[C+1,13]=M1/t.1; Res[C+1,14]=ybar2;
             Res[C+1,15]=M2/t.2; Res[C+1,16]=ybar3; Res[C+1,17]=M3/t.3};
}
cs=ce+1;
\end{verbatim}
for(i in 2:24) {
  MINP = min(Res[i,2], Res[i,4], Res[i,11]);
  if(Res[i,2]==MINP) Res[i,18]=0 else {
    if(Res[i,4]==MINP)
      Res[i,18]=1 else Res[i,18]=2;
  }
  HeadingsRes<-t(c("Chromosome","CPIC(n)","j","CPIC(j)","mu1","lambda1",
                  "mu2","lambda2","j1","j2","CPIC(j1,j2)","mu1","lambda1","mu2","lambda2",
                  "mu3","lambda3","ModelSelected")); Res[1,]-HeadingsRes;

  #use write.matrix to create an output file#
A.3 Chromosome Plots

Figure A.1: GM01535 Chromosome 5

Figure A.2: GM01535 Chromosome 12

Figure A.3: GM01750 Chromosome 14
Figure A.4: GM13330 Chromosome 1

Figure A.5: GM13330 Chromosome 4

Figure A.6: GM13031 Chromosome 17
Bibliography


