DETECTING CHANGE–POINTS IN A COMPOUND POISSON PROCESS

A DISSERTATION IN
Mathematics
and
Economics

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

DOCTOR OF PHILOSOPHY

by
Paul J. Plummer

B.S.E., University of Central Missouri, 1990
M.S., University of Central Missouri, 1995

Kansas City, Missouri
2012

DETECTING CHANGE–POINTS IN A COMPOUND POISSON PROCESS

Paul J. Plummer, Candidate for the Doctor of Philosophy Degree

University of Missouri-Kansas City, 2012

ABSTRACT

A Statistical change point problem was first studied in the mid-1950s in the context of quality control in industrial processes. A change point is defined as a point in the time order when the probability distribution of a sequence of observations differs before and after that point. The literature of statistical change point has evolved over time and now includes a significant amount of scholarly work on change point analysis with many important applications in other disciplines such as economics, geosciences, medicine, and genetics, to name a few.

This work examines the problem of locating changes in the distribution of a Compound Poisson Process where the variables being summed are iid normal and the number of variable follows Poisson. The maximum likelihood ratio for the location of the change point will be explored as well as an information criterion developed, for the case of known variance, while a Bayesian approach is used to deal with the case including change in variance. These results can be applied in any field of study where an interest in locating changes not only in the parameter of a normally distributed data set but also in the rate of their occurrence. It has direct application to the

study of gene expression data in cancer research, where it is known that the distances between the genes can affect their expression level.

The faculty listed below, appointed by the Dean of the School of Graduate Studies, have examined a dissertation titled "Detecting Change–Points in A Compound Poisson Process," presented by Paul J. Plummer, candidate for the Doctor of Philosophy degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Jie Chen, Ph.D., Committee Chair
Department of Mathematics and Statistics

Yong Zeng, Ph.D.
Department of Mathematics and Statistics

Hristo Voulov, Ph.D.
Department of Mathematics and Statistics

James I. Sturgeon, Ph.D.
Department of Economics

Karen Vorst, Ph.D.
Department of Economics

CONTENTS

ILLUSTRATIONS

TABLES

ACKNOWLEDGMENTS

A work of this magnitude cannot be accomplished without the help and support of many people. First and foremost, I would like to thank my research advisor, Dr. Jie Chen, for her guidance, patience, and gentle nudging. Dr. Chen is a great teacher, she has inspired and challenged me to be more than I thought I could be. She has shown all the characteristics a person would hope for in a mentor, she has challenged me, cheered me on, and kept me on track to complete my degree.

I would also like to recognize my committee members, Dr. James Sturgeon, Dr. Karen Vorst, Dr. Hiristo Voulov, and Dr. Yong Zeng. I thank them for their support and allowing me the extra time to finish. I know that all of them are busy and I appreciate them taking the time to work with me throughout this process. They have all been very supportive of my research and I immensely appreciate it!

Next, I would like to thank my family for their unending love and support. They have believed in me even when I doubted myself and it is in a large part their sacrifice that has made this work possible. I also thank the Father of my Lord and Savior, Jesus Christ, for enabling me with the ability to complete this dissertation.

Others who have made this work possible include the past and present chair of the Department of Mathematics and Computer Science at the University of Central Missouri, Dr. Ed Davenport and Dr. Xiaodong Yue, for working with my schedule so that I could take courses and complete this degree. Also, I would like to thank the Office of Graduate Studies, for the many answered questions and assistance in

completing this process.

Finally, I would like to thank Dr. Larry Eifler, Professor Emeritus of the Department of Mathematics and Statistics, for providing the LaTeX template used to type this dissertation. This would have been a much greater challenge had I not had the template to use.

DEDICATION

To God, the father, and my Lord, Jesus Christ, to whom all the work of my hands are meant to bring glory. Also to my family, to Cheryl, my loving and devoted wife, my best friend, and my greatest cheerleader, and my two sons, Troy and Cole, for all the times their dad wasn't there when they wanted him to be. I love you all and look forward to our time together in the future.

> But if serving the Lord seems undesirable to you, then choose for yourselves this day whom you will serve,. . ..
> But for me and my house, we will serve the Lord.
>
> Joshua 24: 15

CHAPTER 1

INTRODUCTION

Motivation

In recent years, there has been considerable effort put into studying the genetic structure of cancer versus normal cells. Researchers are aware that some of the causes of cancer are mutations in certain genes [28]. These can be the lack of production (deletions) or the increased production (amplifications). The more information researchers have about the location of the amplifications or deletions, called breakpoints, along the genome, the greater the chance they have for developing treatments and identifying the aggressiveness of a cancer [12]. The identification of these breakpoints is thus essential to the further understanding and treatment of cancer.

To measure these amplifications or deletions microarray comparative genomic hybridization (aCGH) is used. These intensities are denoted as R/G and a ratio of one means there is no change in behavior. By taking $log_2 R/G$, zero is now considered no change and positives represents amplifications while negatives represent deletions. It should also be noted that the aCGH process creates noise which makes it difficult to locate changes and makes it necessary to use a statistical tool like statistical change point analysis [28].

Several different approaches have been applied to this problem including clustering proposed by Segal and Wiemels (2002) [168],which focused on locating Translocation Breakpoints, and Xing et al. (2007) [215] which proposed the use of hierarchical

clustering to locate copy number variations. Smoothing techniques have also been proposed by Jong et al. (2004) [99] using an algorithm called aCGH-smooth, while Bilke et al. (2005) [12] develop a novel algorithm for eliminating systematic noise, and Hsu (2005) [82] used wavelets. A number of software packages have been developed for the analysis of breakpoints including one proposed by Daruwala et al. (2004) [40] which used dynamic programming, ChARM by Myers et al. (2004) [133] which used an expectation-maximization algorithm, CGHAnalyzer by Margolin et al. (2005) [128] which offered a number of methods, ChromoScan by Sun et al. (2006) [185] used a scan statistic, GEAR by Kim et al. (2008) [102] which used functional enrichment analysis using a prior selected functional gene set and allows for user defined cut-off values for identification of amplifications and deletions, and CNVDetector by Chen, Liu, and Chao (2008)[32] which can be used even when the noise is not normally distributed and provides a measure of the statistical significance. Huang et al. (2005) [83] proposed the Lasso Method based on a penalized least squares regression, while Goa and Huang (2010) [53] proposed a LAD regression model with adaptive fused lasso penalty. The use of Hidden Markov models was proposed by Shah et al. (2006) [172] using a robust HMM model while Stjernqvist et al. (2007) [184] used a continuous-index HMM model. Brot and Richardson (2006) [14] proposed the use of a spatially correlated mixture model with three states unmodified, deleted, or amplified. Picard et al. (2005) [154] used a penalized Information Criterion. A circular binary segmentation approached was developed by Olshen and Venkatraman (2004) [141], by splicing the two ends of the data to form a circle, they introduced a modification of the binary segmentation process suggested by Vostrikova (1981) [193],

called circular binary segmentation (CBS), they also suggested the use of a change point model based on a likelihood test but used a permutation method to calculate the p-value. Change point models were also used by Minin et al. (2005) [130] suggesting a dual multiple change-point model, Chen and Wang (2006) [28] applied a Gaussian mean and variance change model (MVCM) using an information criterion to select the best model, while Liu et al. (2008) [122] used a Bayesian change point algorithm. Chen and Wang (2009) [29] did a comparison study to test their MVCM method with Olshen's CBS method, while Koike et al. (2011)[107], compared five widely used packages.

Levin et al. (2005) [117] noted that previous methods treated the data as equally spaced and did not take into account the variation in the distances of the genes. They focus on identifying either potential amplifications or deletions by the following a two step method. First, they assumed that multiple samples have been taken and that each expression level can be normalized using the sample mean and variance for that location over the multiple samples; then they set a threshold and cluster the genes into similar expression levels. Using the distances between the genes and the number of genes sampled in a given distance as a Poison Process, they create a Compound Poisson Process, where genes above the threshold occur at one rate while genes below the threshold occur at a different rate. They then looked for a grouping of genes from the above threshold group that occur at a greater than expected rate. Once a potential cluster of genes was identified, they used a gamma distribution to determine the significance of the clustering. Li and Zhu (2007) [119] also tried to incorporate distances between genes with the use of Fused Quantile Regression. Just

recently, Chen et al. (2010) [30] used a Compound Poisson Process to model the gene expression levels and gene position, by applying a Bayesian approach with a uniform prior for the location of the change.

The distances between the genes play a role in their expression level and should logically be included in the identification process [117]; the cost of attaining the expression levels limits the number of repetitions that may be performed making estimates of the mean and variance for the normalization process impossible to attain or unreliable. One way of incorporating the distances into the process is by viewing the expression data as a Compound Poisson Process, where the $log_2 R/G$ is assumed normally distributed, and then testing for changes in any of the parameters. As will be observed in chapter 2, at present identification of change points in Compound Poisson Processes is an understudied problem.

While the specific application that motivated the focus of this model is from cancer research, change point analysis has been applied to many application areas including economics. Bryden and Carlson (1994) [17] used change point analysis to study disinflation since 1982. Changes in stock prices were studied by both Inclán (1993) [91] and Chen and Gupta (1997) [24]. Freeman (2005) [50] studied changes in the growth effects of state banking deregulation using a change point model. Frisén (2009) [51] discussed the need and use of sequential surveillance, to identify changes points, in many areas including the monitoring of economic data. Also, in the second edition of *Parametric Statistical Change Point Analysis*, Chen, and Gupta (2012) [27], applications to genetics, medicine, and finance are presented. Thus, the model(s) developed in this research could be used to evaluate changes in data from any field

where the variable being measured is assumed normally distributed and there is an interest in identifying changes in not only the parameters of the distribution but also the rate at which they occur.

Outline of Remaining Chapters

The remainder of this work is organized as follows, in Chapter 2 a brief history of existing change point analysis methods will be presented. In Chapter 3, a formal presentation of the problem will be given as well as outline of the methods to be used to address the problem. Chapter 4 includes the development of the Likelihood Method and an Information approach for the case of known variance of the normal variables. In Chapter 5 a Bayesian approach is developed to include changes in the variance of the normal variables. Finally, Chapter 6 discusses open problems and future research goals.

CHAPTER 2

REVIEW OF CHANGE POINT LITERATURE

Introduction to Change Point Analysis

The field of change point analysis began when, Page, published his landmark articles in 1954 [144] and 1955 [145]. Since then the study of change point analysis has grown into two distinct areas of study. The first is known as fixed set or off-line change point analysis. In this area, the data $x_1, x_2, ..., x_n$ is a pre-observed set of measurements, in the order they were observed. The goal is to identify the index(es) at which the distributions of the x_is have changed. A comprehensive look at parametric forms of change point analysis can be found in Chen, and Gupta (2000,2012), [26][27].

The second area of change point analysis is known as sequential or on-line change point analysis. This area deals with an ongoing process in which repeated measurements are taken. Examples would be a manufacturing process or monitoring of a person's hear rate. Like off-line change point analysis the objective is to locate a change in distribution, often times mean. The emphasis here though is identifying the change as soon after it occurs as possible.

The rest of the chapter is organized as follows. Methods that have been developed for fixed set change point analysis will be presented first, looking at both parametric and nonparametric methods. Then a brief review of sequential methods will be given.

A Review of Fixed Set Change Point Methods

Parametric Methods for Change in the Distribution of a Sequence of Normal Observations

One of the first and most heavily studied problems is the location of a change in the mean of a normal distribution. It was first given as an example in Page (1957) [146] using a CUSUM method. Chernoff and Zacks (1964)[35] presented a Bayesian approach to the problem, as well as estimating the current mean. Their work was extended by Gardner (1969) [54]. The asymptotic distribution for the maximum likelihood estimate was derived in Hinkley (1970) [74] as well as the asymptotic distribution for the likelihood ratio test, but an estimate was used in the calculations of the values. Then Sen and Srivastava (1975) [170] and [171] did a comparison between the powers of the Bayesian statistics and the Maximum Likelihood ratio statistics for different alternative hypothesis. They also derived exact and asymptotic distributions for some of the Bayesian statistics.

It was not until Hawkins (1977) [72] that the distribution of the Likelihood Ratio Test Statistic was derived as well as its asymptotic distribution. Like all previous work mentioned, these results assumed that the variance of the distribution was known. His article was also significant because he included results for the case of a fixed variance that was unknown. These results were achieved by the use of a recursive formula. Hawkins' work for the case of unknown variance was later corrected by Worsley (1979) [208], in which he found the distribution and asymptotic distribution by other means.

Because of the complexity of the problem of multiple change points, the pre-

vious works focused on the assumption that at most one change had occurred. It was not until Vostrikova (1981) [193] showed the consistency of the binary segmentation process that multiple change points were considered. Under binary segmentation, the entire data set is analyzed for a single change point. If one is found and is determined significant then the data is split into two sets and each is analyzed separately. This process is continued as long as a significant change point is located in one of the current segments. With binary segmentation allowing existing methods to be used in the case of multiple change points, the focus of research shifted to more complex cases of change in distributions as well as the change in the structure of other distributions.

However, as the use of change point in more fields has accelerated in the last few years, the case of a change in the mean of a normal distribution has reemerged. Sofronov (2001) [179] and (2005) [180] construct an asymptotically d-optimal test of a change point. Fotopoulos, and Jandhyala (2007) [48], refine the estimates used by Hinkley, while Zhang and Siegmund (2007) [220] a modified BIC was introduced for the detection of the number and location of the change points. Wang et al. (2008) [195] used local averages or "localized information" to detect changes in mean. Erdman and Emerson (2008) [45] proposed a fast Bayesian change point method for detecting multiple change points. Chen, Cohen, and Sackrowitz (2011)[31], showed the consistency of the binary segmentation and maximum residual down methods for locating multiple change points.

Another problem involving the normal model is finding the location for a shift in variance assuming a fixed mean. This problem was studied by Hsu (1977) [81] under the assumption that the initial value of the variance was known. He developed

two tests one that was locally most powerful and another based on the CUSUMs of Chi Squared values. The problem was also considered by Inclán (1993) [91], who developed a Bayesian procedure for detecting multiple change points using posterior odds, while Inclán and Tiao (1994) [92] proposed using and iterative cumulative sum of squares algorithm to locate multiple changes. Use of the Schwarz information criterion (SIC) was proposed by Chen and Gupta (1997) [24] for the location of a single change point and the asymptotic null distribution of their statistic is presented.

The case for change in both mean and variance of a set of normal observations was first considered by Horvath (1993) [76]. Horvath developed the asymptotic distribution of the likelihood ratio statistic. Chen and Gupta (1999) [25] propose a test based on the (SIC) and develop its asymptotic distribution. The use of a product partition model (PPM) is proposed for identification of multiple change points first by Barry and Hartigan (1992) [9], and again by Loschi and Cruz (2002) [124],[125] and (2005) [126]. An information criterion for testing a single change in parameter is developed by Chen et al. (2006) [22]. Then a modified information criterion, MIC, is then developed by Pan and Chen (2006) [147] for testing multiple change points, where the number of change points is known. The penalty is based on the distance between change points where a model that evenly splits the data set is given preference. While the normal model is one of the examples given, the MIC as with all information criterion may be applied to any regular model.

Other work involving normal models include Menzefricke (1981) [129] who looked at changes in the precision of normal random variables, which is the inverse of the variance. In his work he assumed both changes in precision and mean. He

developed the posterior distribution for the location of a single change as well as the magnitude of the change. Tourneret et al. (2004) [190] derived the Cramer-Rao lower bounds for estimators of changes in models with additive and multiplicative noise. Lavielle (2005) [114], developed a method using penalized contrasts which may be applied to even non-Gaussian sequences but estimated the change using a Gaussian log likelihood function.

Parametric Methods for Change in the Distribution of a Sequence of Multivariate Normal Observations

Like the univariate normal model, early work with sequences for multivariate normal models focused on changes in the means only. The first to consider the problem was Sen and Srivastava (1973) [169], who assumed at most one change and proposed a test statistic based on a Bayesian approach as well as its exact and asymptotic distribution. A likelihood ratio test statistic for locating a single change was then introduced by Srivastava and Worsley (1986) [182] and gave an approximation for the null distribution. Krishnaiah et al. (1990) [110] suggested the use of a local likelihood method for the case of multiple changes. James et al. (1992) [93] gave an approximation for the asymptotic distribution of the likelihood ratio test. An information approach was presented by Gupta and Chen (1996) [61] using the SIC and presented applications in geology and literature. Rukhin (2002) [167] studied the asymptotic behavior of posterior distributions of the change point parameter. For the case of a two dimensional space, Ninomiya (2004) [135] developed an improved approximation for the likelihood ratio test. Ninomiya (2005) [136] studied the properties of AIC-type information criterion for the multivariate change point model. Son and

Kim (2005) [181] presented a Bayesian single change point method for the detection of change in mean, variance, or mean and variance. Gooijer (2006) [43] developed a test for change in the sequence of covariance matrices and presented the tests asymptotic distribution under the null hypothesis. Lastly, Fotopoulos et al. (2009) [49] studied the asymptotic properties of the MLE for change in both mean vector and covariance matrix under different alternative hypothesis.

Parametric Methods for Change in a Regression Model

Regression analysis plays an important role in many real world applications. This makes refinement of regression models critical to the advancement of Statistics. Part of this refinement is identifying when a model has had a change in parameter. The problem of change in the parameter of a regression model has been studied by many authors since Quandt (1958) [157] and Quandt (1960) [158], where the likelihood ratio test for identifying the location of the test and estimating the parameters were explored. Later, Ferreira (1975) [46] used a Bayesian analysis for a switching regression model, where the number of changes is known. A Bayesian approach was also used by Chin Choy and Broemeling (1980) [36] for a changing linear model and by Smith and Cook (1980) [177] who applied their model to renal transplant data. An information approach is proposed by Chen (1998) [23] who suggested the use of the SIC for both simple and multiple linear regression models. A method to estimate least squares regression models with multiple changes at unknown points is given by Bai and Perron (1998) [8], while Charlton and Troskie (1999) [19] used a Bayesian approach for the case of multiple regression models with autocorrelated errors. The aymptotics of the maximum likelihood estimator for the case of at most one change

was studied by Koul and Qian (2002) [109]. Bischoff et al. (2004) [13] studied the Kolmogorov test for trends of a Brownian bridge and applied their results to the change point problem in regression models. A set of generalized maximum likelihood asymptotic power one test for different cases of change in a linear regression model was given by Gurevich and Vexler (2006) [64]. Wu (2008) [213] used an information criterion approach to simultaneously locate change points and select variables for multiple linear regression models.

Other regression models studied include Pons (2002) [156] which considers a Cox regression model with at most one change. Chiu et al. (2005) [37] suggested the use of bent-cable regression using the least squares estimates. A MLE approach was proposed by Gurevich and Vexler (2005) [63] for locating change points in logistic regression models. The measurement error model was studied by Quintana and Iglesias (2005) [159] using a Bayesian approach. Dupuy (2006) [44] studied the problem of estimating the location of change points in hazard regression models. Nosek and Szkutnik (2008) [139] studied changes in models of variables that are linearly related to time. Detection of changes in growth processes was proposed by Ninomiya and Yoshimoto (2008) [137]. Ciuperca (2009) [38] studied the properties of M-estimation in the case of multiple change points in nonlinear models.

The basic assumption of most change point models is that a change occurs abruptly at a fixed point. In many real world applications, it might be more reasonable to assume that a change occurs gradually over time. Gill and Baron (2004) [55] considered the case of a gradual change in the distribution of a general regression model. They constructed the maximum likelihood estimator for the change and

showed that it is consistent.

Parametric Methods for Change in a Time Series Model

The use of stationary time series models in the study of the financial sector, as well as other applied areas, makes identification of change points in these models important. While several models have been considered, most attention has been placed on the autoregressive model beginning with Wichern et al. (1976) [205] who studied changes in variance by a two-step method. Other researchers that focused on the autoregressive model include Picard (1985) [153] who developed a test for the change in the parameters, Takeuchi and Yamanishi (2006) [188] who presented a framework for identifying changes in distribution and outliers simultaneously, Olsen at al. (2008) [142] who focused on changes in mean, variance, and first lag parameter, and Gombay (2008) [57] who assumed Gaussian white noise and looked for changes in mean, variance, and all lag parameters. Gombay also considered the case of a temporary change where after a time the distribution returned to the original one.

Other models studied include the fractionally integrated ARMA model with Gaussian white noise by Ray and Tsay (2002) [163] who applied a Bayesian approach to changes in mean and the dependency variables. Hariz and Wylie (2005) [70] used a CUSUM test to identify change in mean of a stationary process with long-range dependence. Wang and Wang (2006) [196] considered changes in the dependency parameter in moving average model. The moving average model was also studied by Wang (2008) [198] for a change in mean. Reboul and Benjelloun (2005) [164] proposed a Bayesian fusion approach to change points in a piecewise stationary process, while Last and Shumway (2008) studied the problem of identifying changes in the power

spectrum of piecewise locally stationary time series and showed the consistency of their process. Finally, Davis et al. (2008) [41] considered the case of a multiple changes in parameter of nonlinear time series models and developed a general algorithm. Jirak (2012) [98], studied changes in time series where the dimension of the parameter space may increase with the sample size.

As with most change point problems the above studies focused on abrupt changes. An exception to that was Wang (2007) [197] who focused on identifying gradual change in mean in a long memory process. Along with presenting a test statistic, they also found its limit distribution and proved its rate of consistency.

Parametric Methods for Change in Non-Gaussian Models

Besides the normal distribution, there are a few other well known distributions which have been studied for change point detection. Changes in a sequence of binomial random variables was first studied by, Hinkley and Hinkley (1970) [75], they found an approximation to the asymptotic distribution of the likelihood ratio test statistic for a single change point. A Bayesian Approach was presented in Smith (1975) [176], while the CUSUM statistic was presented by Pettitt (1980) [152] and its asymptotic properties explored. Worsly (1983) [209] did a study of the powers of both the likelihood ratio and CUSUM tests and determined that the likelihood ratio is more powerful when the change occurs near the beginning or the end of the data. A test for multiple changes in a binomial distribution was presented in Fu and Curnow (1990) [52] using maximum likelihood estimation, Stephens (1994) [183] using a Bayesian approach, and Albert et al. (2004) [3] in which a penalized likelihood ratio test was used to determine the number of change points and the location.

The exponential model is another model which has been studied. Worsley (1986) [210] explored confidence regions and tests for change in an exponential family of random variables, which included the exponential variable, using the likelihood method. Worsley's article was followed by Haccou et al. (1988) [67], who derived the likelihood ratio test for exponentially distributed random variable for a single change. Hoccou and Meelis (1988) [66] considered a test that compares the possibility of $n-1$ change points to 2 change points, then 2 change points to 1 change point, and then 1 change point to no change points. Gupta, and Ramanayake (2001) [62] studied changes in the exponential distribution that follow a linear trend rather than an abrupt change. Karasoy and Kadilar (2007) [100] presented a Bayesian methods, while Karasoy and Kadilar (2009) [101] presented two alternative methods for estimating the change point. A related article was Víšek (2003) [192] in which the likelihood ratio test statistic for the double exponential model was derived and its asymptotic properties explored.

Another distribution used in a wide range of applications is the Poisson distribution. The Poisson distribution has been studied from a Bayesian approach by Raftery, and Akman (1986) [162], and Carlin et al. (1992) [18] for at most one change, while Loschi and Cruz (2005) [127], and Tian et al. (2009) [189] considered the case of multiple changes. Watkins and Yang (2005) [204] used a generalized likelihood ratio test to identify a single change point and multiple change points were located by using the method recursively.

Other distributions that have been studied include the Gamma and Chi-Squared by Jandhuala et al. (2002) who derived a general likelihood ratio test for

change in variability of a process. Lin (2008) [120] presented a two-stage failure model for Bayesian change point analysis and applied it to the gamma distribution. The Bivariate Gamma distribution by Chatelain et al. (2007) [20], and Chatelain et al. (2008) [21] who used a test based on the maximum likelihood principle. Hušková and Neuhaus (2004) [89] developed a method for working with censored data, and Fotopoulos (2009) [47] studied the convergence rates of change point estimators in general. Ng (2008) [138] used an EM algorithm to estimate the maximum likelihoods estimators efficiency, the method was then demonstrated using a compound poisson process, where the variables being summed were exponentially distributed.

Williams and Kim (2011) [207], developed a likelihood ratio test for a continuous monotone hazard function. Iacus and Yoshida (2012) [90] proposed a quasi-maximum likelihood method for identifying the location of change points in the volatility of a stochastic differential equation.

Shen and Ghosh (2011) [173] developed a new analogue l BIC method for exponential family of distributions and showed its approximation error had the same rate of convergence as the Schwartz BIC. Lai, and Xing (2011) [112] constructed a Bayesian approach for locating multiple change points in multi parameter exponential family distributions.

Nonparametric Methods for Locating Changes in Fixed Set Data

As with the parametric methods, the earliest work in nonparametric work in change point analysis looked for changes in mean of normally distributed variables. Bhattacharyya and Johnson (1968) [11] developed a locally best invariant test, assuming the initial mean was known and applied it to the normal distribution as well

as the double exponential. Others who considered the problem of change in mean were, Račkauskas and Suquet (2004) [160] who proposed a Hölder norm test statistic for the change in mean of Non-Gaussian iid variables, as well as Orasch and Pouliot (2004) [143] who used Weighted Sup-Norm Functions to locate changes in mean of a distribution with variance equal to one. Wang et al. (2008) [203] developed a non-parametric test based on the difference in sample medians, and finally Cheng (2009) [34] created an efficient algorithm for estimating change in mean, which can also be adapted to locate changes in variance with constant mean. Oh et al. (2005) [140] suggested the use of artificial neural networks to identify changes in variance. Methods for finding changes in mean and variance together were presented in Müller and Wai (2006) [132] using fluctuations of mutagrams, while Kirch (2007) [103] proposes a block permutation method for the case of at most one change. A minimally selected p method for locating single changes in a Binomial Sequence was proposed by Halpern (1999) [68]. Later a method for locating multiple change points in Binomial Sequence was proposed in Halpern (2000) [69]. A Nelson–Aalen type estimator was proposed by Wu et al. (2003) [211] for locating changes in a hazard function.

The following articles deal with changes in non specific distributions. Gombay and Horváth (2002) [58] studied the convergence rates of U-statistic processes and their bootstapped versions. Horváth and Hušková (2005) [78] used permutations of U-statistics to locate change in distribution. Li and Lin (2007) [118] investigated the location of change in density function where the random sequence was assumed to be associated using functionals. Horváth and Shao (2007) [79] studied the limit behavior of permutations of empirical processes and applied it to change point analysis. Finally,

Antoch et al. (2008) proposed the use of a data driven rank test.

Regression models have also been studied by many authors. Müller (1992) [131] developed a method for locating location and size of change points in regression models by comparing one-sided kernel smoothers. Loader (1996) [123] proposed the use of a one sided nonparametric regression estimates of the mean function, then Horváth and Kokoszka (2002) [77] expanded upon Loader's method by using local polynomial smoothing and showed the consistency of their method is also demonstrated. Grégoire and Hamrouni (2002) [59] developed two nonparametric tests based on local smoothing, using kernel smoothing, while Grégoire and Hamrouni (2002) [60] used local linear smoothing and applied their method to a compound Poisson process with drift. Local polynomial fits were proposed by Huh and Carrière (2002) [84] to locate changes in a derivative of a regression function, Huh and Park (2004) [85] for models when design points were random, and Lin et al. for locating changes in derivatives or the function itself. Yu et al. (2005) [216] developed a method for locating change points by using Quadratic Programming. Kernel methods are proposed by Cheng and Raimondo (2008) [33] for the location of changes in the first derivative of a regression model. Fuzzy multiple objective programming was proposed by Yu and Tzeng (2009) [217] to find the number and location of change points.

As stated earlier, identification of change points in Time Series models was an important area of research. There have been several researchers who proposed nonparametric tests for this purpose including, Rozenholc (2001) [166] who developed two Kolmogorov-Smirnov type statistics for the identification of changes in the spectral characteristics of Gaussian tapered data. Kumar and Wu (2001) [111] developed

a test based on fuzzy statistics to identify gradual changes in time series. Hušková (2003) [86] then developed a serial rank statistic to identify a change in the independence of a data set. Lee and Na (2004) [115] proposed a test based on the sequential kernel estimate to locate changes in the marginal density function of a strong mixing process. The problem of identifying changes in the marginal density function of a dependent sequence is then studied by Hariz et al. (2005) [71] who used empirical measures and semi-norms. Lirch and Steinebach (2006) [105] studied changes, both abrupt and gradual, in the mean and variance of a strong invariance data set. Confidence intervals for change points in time series are then developed using a block bootstrapping by Hušková and Kirch (2008) [87]. Bootstrapping was also used by Hušková et al. (2008) [88] to detect changes in autoregressive time series, Nie et al. (2009) [134] developed a nonparametric test for changes near the ends of long-range dependent sequences. Finally, Wang (2008) [199] used a Wilcoxon-type rank statistic to detect changes in long-memory time series models.

Wavelet Methods for Locating Changes in Fixed Set Data

Since the development of wavelets, the applications in both mathematics and statistics has been ever increasing. Because of their ability to break data into different band widths, they are useful in the location of breaks in local data. Wang (1995) [201] introduced a wavelet method for the detection of jumps in data with the presence of white noise. Wang (1999) [202] suggested the use of wavelets for indirect data. Antoniadis and Gubels (2002) [4] proposed the use of wavelets in determining the number and location of changes in a regression model with white Gaussian noise. Park and Kim (2004) [148] used boundary wavelets to locate a sharp change

point in a regression function. Ko and Vannucci (2006) [108] developed a Bayesian wavelet-based method for locating multiple change points in long memory parameters of various time series models. Jansen (2007) [95] used a wavelet method to locate multiscale changes in a Poisson count data. Plonka and Ma (2007) [155] studied the convergence rate of wavelets for denoising piecewise constant images while preserving discontinuities. Wang and Cai (2010) [200] developed a wavelet method for identifying change points in random design regression models with long memory errors.

A Review of Sequential Change Point Methods

Since the purpose of sequential change point analysis is the continual monitoring of ongoing systems, a vast majority of the literature focuses on the location of a single change. One of the key elements in the development of sequential sets for production schemes is the average run length (ARL). The ARL is a measure of the cost of making a wrong decision, either the cost of stopping production when the system has not broken down, or the cost of producing poor quality, unusable goods [144]. Page (1954) [144] introduced the idea of a cumulative sum or CUSUM statistic for identifying a change in the mean of a normal sequence with fixed variance. Goel and Wu (1971) [56] studied the ARL of CUSUM charts for normal mean. A related article is Aue and Horváth (2004) [7] which studied the delay time, the time or number of observations after a change it takes to determine the change has occurred in a weighted CUSUM detection scheme. Wu (2005) [212] proposes a CUSUM procedure for identifying changes in both mean and variance of a normal sequence. Wu et al. (2009) [214] proposed an adaptive CUSUM to locate changes in mean of a normally distributed sequence. Jin et al. (2009) [97] extended the use of a CUSUM

type statistics to locate changes in the mean of heavy-tailed distributions. Shi et al. (2009) [174] studied the convergence rate of CUSUM type statistics.

An alternative to the CUSUM control charts was given in Roberts (1959) [165] based on geometric moving averages; this led to the exponentially weighted moving average or EWMA control charts. Crowder (1987) [39] proposed a simple method for calculating the ARL of EWMA charts. Knoth (2003) studied the performance of EWMA charts for change in the mean of normal observations, where the control limits were allowed to vary over time. Gut and Steinebach (2004) [65] studied EWMA charts for changes in the mean of a stochastic process. Jiang et al. (2008) [96] created an adaptive CUSUM procedure which includes an EWMA-based shift estimator.

Other methods for identifying changes in mean of iid random sequences include, a nonlinear filtering approach in Vellekoop and Clark (2006) [191], a bootstrapping method in Kirch (2008) [104], a closed-end fluctuation test in Horváth et al. (2008) [80], a binary control chart for detecting small jumps in Rafajlowicz and Steland (2009) [161], and using a hybrid fuzzy-statistical clustering method in Alaeddini et al. (2009) [2]. A method for identifying multiple changes in mean was proposed in Arunajadai (2009) [6] using a robust resistant statistical procedure.

Changes in the Poisson Process is another problem which has received much attention from authors. In the sequential setting, the Poisson Process was studied in Zacks (1991) [218] which focused on the stopping times for Poisson Processes with linear boundaries. The optimal stopping time was then explored by both Herberts and Jensen (2004) [73], and Brown and Zacks (2006) [15]. Zacks (2004) [219] determined the exact distribution of a one-sided CUSUM procedure for a Poisson Procedure.

Perry et al. (2007) [149] developed a control chart for monitoring step changes in the rate parameter of a Poisson Process which adjusts to the magnitude of the change. A Bayesian method for identifying change when the arrival rates are unknown was presented in Brown (2008) [16]. Bayraktar and Sezer (2009) [10] also used a Bayesian method but assumed the prior distribution of the arrival time was known and used a phase-type prior distribution.

CHAPTER 3

STATEMENT OF THE PROBLEM

The Problem

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

Let [0,T] be divided into ℓ non-overlapping intervals each of length T_j. Then in each interval there are m_j occurrences and $m_j \sim POI(\lambda T_j)$. Now let, $M_j = \sum_{i=1}^{j} m_j$ and $M_0 = 0$. So, $X_{t_j} = \sum_{i=M_{j-1}+1}^{M_j} Y_i$, for $j = 1, 2, \cdots, \ell$. Then $X_{t_1}, X_{t_2}, ..., X_{t_\ell}$ forms a Compound Poisson Process dependent on (λ, μ, σ^2), and where, given m_j, $X_{t_j} \sim N(m_j\mu, m_j\sigma^2)$ and $X_{t_1}, X_{t_2}, ..., X_{t_\ell}$ are independent.

The change point problem can then be expressed as testing the following null hypothesis:

$H_0 : X_{t_1}, X_{t_2}, ..., X_{t_\ell}$ form a homogenous Compound Poisson Process depending on $(\lambda_0, \mu_0, \sigma_0^2)$

versus the alternative hypothesis

$H_1 : X_{t_1}, X_{t_2}, ..., X_{t_{j_1}}$ depend on $(\lambda_1, \mu_1, \sigma_1^2)$, $X_{t_{j_1}+1}, X_{t_{j_1}+2}, ..., X_{t_{j_2}}$ depend on $(\lambda_2, \mu_2, \sigma_2^2)$, $\cdots, X_{j_{t_p}+1}, X_{j_{t_p}+2}, ..., X_{t_\ell}$ depend on $(\lambda_p, \mu_p, \sigma_p^2)$ with $(\lambda_1, \mu_1, \sigma_1^2) \neq (\lambda_2, \mu_2, \sigma_2^2) \neq \cdots \neq (\lambda_{p+1}, \mu_{p+1}, \sigma_{p+1}^2)$.

Before developing a test for this multiple change point problem, the following

single change problem will be explored.

$$H_0 : X_{t_1}, X_{t_2}, ..., X_{t_\ell} \text{ form a homogenous Compound Poisson Process depending on } (\lambda_0, \mu_0, \sigma_0^2) \tag{3.1}$$

versus the alternative hypothesis

$$H_1 : X_{t_1}, X_{t_2}, ..., X_{t_j} \text{ depend on } (\lambda_1, \mu_1, \sigma_1^2) \text{ while } X_{t_j+1}, X_{t_j+2}, ..., X_{t_\ell} \text{ depend on } (\lambda_2, \mu_l, \sigma_2^2) \text{ with } (\lambda_1, \mu_1, \sigma_1^2) \neq (\lambda_2, \mu_2, \sigma_2^2). \tag{3.2}$$

The results will then be extended to the case of multiple changes.

Methods to be Used

For the case of known variance of the normal variables, a likelihood ratio procedure and information approach will be explored. The consistency of the likelihood estimators will be proved. For the case including unknown and possible changing variance in the normal variables, a Bayesian method is developed. Each method is extensively tested using simulations and then each is applied to the analysis of the aCGH data.

CHAPTER 4
CHANGE POINTS IN COMPOUND POISSON PROCESSES WITH KNOWN VARIANCE

Introduction

The methods developed in this chapter will be under the assumption that there is a known variance σ^2 that is constant through out the sample. The case of change in variance will be discussed in chapter 5 using a Bayesian approach. Unfortunately, without putting additional restrictions on the parameter space of σ^2 the likelihood function of the compound Poisson process with unknown variance does not possess the same desirable properties as the model with known variance.

Likelihood Ratio Procedure Method

As can be observed in the literature review, from Chapter 2, the maximum likelihood ratio procedure plays an essential role in the exploration of many of the statistical change point problems. It is often explored first because of its importance in other estimation techniques such as Information criterion, whose distributions are based on the $-2\log(\Lambda)$ where Λ is the maximum likelihood ratio. To develop the maximum likelihood ratio the likelihood function for both the null and alternative hypothesis must first be found. Then the maximum likelihood estimators, or MLEs, for each of the parameters can be found. By substituting the MLEs into the likelihood functions under the null and alternative hypotheses the likelihood ratio can then be formed.

The Likelihood Function

Let (X, N) form a compound Poisson process of normal variables with mean of μ and variance σ^2. Let λ be the unit rate of the Poisson process and T the length of the interval measured. Then the pdf of (X, N) is as follows:

$$f(x,n) = \begin{cases} \frac{1}{\sqrt{2\pi n\sigma^2}} e^{\frac{(x-n\mu)^2}{2n\sigma^2}} \frac{(T\lambda)^n e^{-T\lambda}}{n!} & \text{if } n \neq 0, \\ e^{-T\lambda} & \text{if } n = 0. \end{cases} \tag{4.1}$$

Let $(x_{t_1}, m_1), (x_{t_2}, m_2), \ldots, (x_{t_l}, m_l)$ be the sample described in Chapter 3, define $A = \{i | m_i \neq 0\}$ and M_A as the cardinal number of A. Then the likelihood function under the null hypothesis (3.1) is given by:

$$\begin{aligned} L_0(\mu, \lambda) &= \left(\prod_{i \in A} \frac{1}{\sqrt{2\pi m_i \sigma^2}} e^{-\frac{(x_{t_i} - m_i\mu)^2}{2m_i\sigma^2}} \right) \left(\prod_{i=1}^{\ell} \frac{(T_i\lambda)^{m_i} e^{-T_i\lambda}}{m_i!} \right) \\ &= (2\pi\sigma^2)^{-\frac{M_A}{2}} \left(\prod_{i \in A} m_i^{-\frac{1}{2}} \right) e^{-\frac{1}{2\sigma^2} \sum_{i \in A} \frac{(x_{t_i} - m_i\mu)^2}{m_i}} \\ &\quad \cdot \left(\prod_{i=1}^{\ell} \frac{T_i^{m_i}}{m_i!} \right) \lambda^{\sum_{i=1}^{\ell} m_i} e^{-\lambda \sum_{i=1}^{\ell} T_i}. \end{aligned} \tag{4.2}$$

For the likelihood function under the alternative hypothesis (3.2), define $A_j = \{i | m_i \neq 0 \text{ and } 1 \leq i \leq j\}$, $A_\ell = \{i | m_i \neq 0 \text{ and } j+1 \leq i \leq \ell\}$, while M_{A_j} and M_{A_ℓ} the cardinal numbers of A_j and A_ℓ respectively. Then the likelihood function is given by:

$$\begin{aligned} L_1(\mu_1, \mu_2, \lambda_1, \lambda_2) &= \left(\prod_{i \in A_j} \frac{1}{\sqrt{2\pi m_i \sigma^2}} e^{-\frac{(x_{t_i} - m_i\mu_1)^2}{2m_i\sigma^2}} \right) \left(\prod_{i=1}^{j} \frac{(T_i\lambda_1)^{m_i} e^{-T_i\lambda_1}}{m_i!} \right) \\ &\quad \cdot \left(\prod_{i \in A_\ell} \frac{1}{\sqrt{2\pi m_i \sigma^2}} e^{-\frac{(x_{t_i} - m_i\mu_2)^2}{2m_i\sigma^2}} \right) \left(\prod_{i=k+1}^{\ell} \frac{(T_i\lambda_2)^{m_i} e^{-T_i\lambda_2}}{m_i!} \right) \end{aligned}$$

$$
\begin{aligned}
= \quad & (2\pi\sigma^2)^{-\frac{M_{A_j}}{2}} \left(\prod_{i\in A_j} m_i^{-\frac{1}{2}}\right) e^{-\frac{1}{2\sigma^2}\sum_{i\in A_j}\frac{(x_{t_i}-m_i\mu_1)^2}{n_i}} \left(\prod_{i=1}^{j}\frac{T_i^{m_i}}{m_i!}\right) \\
& \cdot\lambda_1^{\sum_{i=1}^{j} m_i} e^{-\lambda_1\sum_{i=1}^{j} T_i} (2\pi\sigma^2)^{-\frac{M_{A_\ell}}{2}} \left(\prod_{i\in A_\ell} m_i^{-\frac{1}{2}}\right) \left(\prod_{i=j+1}^{\ell}\frac{T_i^{m_i}}{m_i!}\right) \\
& \cdot e^{-\frac{1}{2\sigma^2}\sum_{i\in A_\ell}\frac{(x_{t_i}-m_i\mu_2)^2}{m_i}} \lambda_2^{\sum_{i=j+1}^{\ell} m_i} e^{-\lambda_2\sum_{i=j+1}^{\ell} T_i} \\
= \quad & (2\pi\sigma^2)^{-\frac{M_A}{2}} \left(\prod_{i\in A} m_i^{-\frac{1}{2}}\right) \\
& \cdot e^{-\frac{1}{2\sigma^2}\left[\sum_{i\in A_j}\frac{(x_{t_i}-m_i\mu_1)^2}{m_i}+\sum_{i\in A_\ell}\frac{(x_{t_i}-m_i\mu_2)^2}{m_i}\right]} \left(\prod_{i=1}^{\ell}\frac{T_i^{m_i}}{m_i!}\right) \\
& \cdot\lambda_1^{\sum_{i=1}^{j} m_i} \lambda_2^{\sum_{i=j+1}^{\ell} m_i} e^{-\lambda_1\sum_{i=1}^{j} T_i} e^{-\lambda_2\sum_{i=j+1}^{\ell} T_i}. \qquad (4.3)
\end{aligned}
$$

The Maximum Likelihood Estimators

Under the null hypothesis, H_0 the log-likelihood function is,

$$
\begin{aligned}
\log(L_0) \quad = \quad & -\frac{M_A}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i\in A}\log(m_i) - \frac{1}{2\sigma^2}\sum_{i\in A}\frac{(x_{t_i}-m_i\mu)^2}{m_i} \\
& + \sum_{i=1}^{\ell} m_i\log(T_i) - \sum_{i=1}^{\ell}\log(m_i!) + \log(\lambda)\sum_{i=1}^{\ell} m_i - \lambda\sum_{i=1}^{\ell} T_i. \qquad (4.4)
\end{aligned}
$$

Denote, $\ell_0 = log(L_0)$ and take the derivatives with respect to μ and λ the following system of patrials are derived:

$$
\frac{\partial\ell_0}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i\in A}(x_{t_i} - m_i\mu) = \frac{1}{\sigma^2}\left[\sum_{i\in A} x_{t_i} - \mu\sum_{i\in A} m_i\right] \qquad (4.5)
$$

and

$$
\frac{\partial\ell_0}{\partial\lambda} = \frac{\sum_{i=1}^{\ell} m_i}{\lambda} - \sum_{i=1}^{\ell} T_i. \qquad (4.6)
$$

Let $\hat{\mu}$ and $\hat{\lambda}$ represent the MLEs for μ and λ respectively. Then setting (4.5) and (4.6) equal to zero and solving, the MLEs are obtained as:

$$\hat{\mu} = \frac{\sum_{i=1}^{\ell} x_{t_i}}{\sum_{i=1}^{\ell} m_i} \tag{4.7}$$

and

$$\hat{\lambda} = \frac{\sum_{i=1}^{\ell} m_i}{\sum_{i=1}^{\ell} T_i}. \tag{4.8}$$

Note: Since $m_i = 0$ and $x_{t_i} = 0$ when $i \notin A$, their sums over $i \in A$ and their sums from $i = 1$ to ℓ is the same, thus simplifying the calculation of $\hat{\mu}$.

Now, under the alternative hypothesis, H_1, the log-likelihood function is given by,

$$\begin{aligned} \log(L_1) \quad = \quad & -\frac{M_A}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i\in A}\log(m_i) - \frac{1}{2\sigma^2}\sum_{i\in A_j}\frac{(x_{t_i} - m_i\mu_1)^2}{m_i} \\ & + \sum_{i=1}^{\ell} m_i \log(T_i) - \sum_{i=1}^{\ell}\log(m_i!) + \log(\lambda_1)\sum_{i=1}^{j} m_i - \lambda_1 \sum_{i=1}^{j} T_i \\ & - \frac{1}{2\sigma^2}\sum_{i\in A_\ell}\frac{(x_{t_i} - m_i\mu_2)^2}{m_i} + \log(\lambda_2)\sum_{i=j+1}^{\ell} m_i - \lambda_2 \sum_{i=j+1}^{\ell} T_i \end{aligned} \tag{4.9}$$

Denote $\ell_1 = log(L_1)$ and take the partial derivatives of ℓ_1 with respect to μ and λ, the following system of patrials are derived:

$$\frac{\partial \ell_1}{\partial \mu_1} = \frac{1}{\sigma^2}\left[\sum_{i\in A_j} x_{t_i} - \mu_1 \sum_{i\in A_j} m_i\right], \tag{4.10}$$

$$\frac{\partial \ell_1}{\partial \mu_2} = \frac{1}{\sigma^2}\left[\sum_{i\in A_\ell} x_{t_i} - \mu_2 \sum_{i\in A_\ell} m_i\right], \tag{4.11}$$

$$\frac{\partial \ell_1}{\partial \lambda_1} = \frac{\sum_{i=1}^{j} m_i}{\lambda_1} - \sum_{i=1}^{j} T_i, \tag{4.12}$$

$$\frac{\partial \ell_1}{\partial \lambda_2} = \frac{\sum_{i=j+1}^{\ell} m_i}{\lambda_2} - \sum_{i=j+1}^{\ell} T_i. \tag{4.13}$$

Let $\hat{\mu_1}$, $\hat{\mu_2}$, $\hat{\lambda_1}$, and $\hat{\lambda_2}$ represent the MLEs for μ_1, μ_2, λ_1 and λ_2 respectively. Then setting equations (4.10) through (4.13) equal to zero and solving, the MLEs are obtained as:

$$\hat{\mu_1} = \frac{\sum_{i=1}^{j} x_{t_i}}{\sum_{i=1}^{j} m_i}, \tag{4.14}$$

$$\hat{\mu_2} = \frac{\sum_{i=j+1}^{\ell} x_{t_i}}{\sum_{i=j+1}^{\ell} m_i}, \tag{4.15}$$

$$\hat{\lambda_1} = \frac{\sum_{i=1}^{j} m_i}{\sum_{i=1}^{j} T_i}, \tag{4.16}$$

and

$$\hat{\lambda_2} = \frac{\sum_{i=j+1}^{\ell} m_i}{\sum_{i=j+1}^{\ell} T_i}. \tag{4.17}$$

Consistency of the Maximum Likelihood Estimators

It's known that under Wald conditions, Wald (1949)[194], the maximum likelihood estimators are consistent. Therefore, we shall first prove a set of properties similar to the Wald conditions for the compound Poisson model. Then we will prove the consistency of the MLE's. As part of this development, the cumulative distribution function will be discussed so it is given below.

Let $\Phi(z)$ be the cumulative distribution function of the standard normal distribution, and (X_{t_i}, N_{t_i}) form the Compound Poisson process described above. Then the cumulative distribution function of (X_{t_i}, N_{t_i}) is,

$$F(x_i, n_i) = \begin{cases} 0 & \text{if} \quad n_i < 0 \\ e^{-t_0\lambda} & \text{if} \quad n_i = 0 \\ e^{-t_0\lambda} + \sum_{j=1}^{n_i} \frac{(t_j\lambda)^j e^{-t_j\lambda}}{j!} \Phi\left(\frac{x-j\mu}{j\sigma}\right) & \text{if} \quad n_i = 1, 2, \cdots \end{cases} \tag{4.18}$$

LEMMA 4.1. Let (X_{t_i}, N_{t_i}) form a Compound Poisson Process with $X_{t_i} \sim N(N_{t_i}\mu, N_{t_i}\sigma^2)$ and $N_{t_i} \sim POI(t_i\lambda)$. Let $f(x, m)$ be the pdf of the process and $F(x, m)$ it's distribution function. Then the following hold:

(i) The probability distribution function $f(x, m)$ is discrete for $m = 0$ and a sequence of absolutely continuous functions with a uniform bound for $m = 1, 2, \cdots, n$.

(ii) The cumulative distribution function $F(x, m)$ is discrete for $m = 0$ and a sequence of absolutely continuous functions for $m = 1, 2, \cdots, n$.

PROOF. Part(i), consider the case for $m = 0$. By the definition of $f(x, m)$,

$$f(y, 0) = \begin{cases} e^{-T\lambda} & \text{if} \quad y = 0 \\ 0 & otherwise. \end{cases} \tag{4.19}$$

Now for $m = 1, 2, \cdots$ let $f(y, m) = f_m(y)$, then

$$f_m(y) = (2\pi m\sigma^2)^{-\frac{1}{2}} e^{-\frac{(y-m\mu)^2}{2m\sigma^2}} \frac{(T\lambda)^m e^{-T\lambda}}{m!}. \tag{4.20}$$

Now to find a uniform bound, consider the first and second derivatives of $f_m(y)$.

$$f'_m(y) = \frac{-2(y - m\mu)}{2m\sigma^2(2\pi m\sigma^2)^{1/2}} e^{-\frac{(y-m\mu)^2}{2m\sigma^2}} \frac{(T\lambda)^m e^{-T\lambda}}{m!} \tag{4.21}$$

and

$$f''_m(y) = \left[\frac{-2}{2m\sigma^2(2\pi m\sigma^2)^{1/2}} + \frac{4(y - m\mu)^2}{4m^2\sigma^4(\pi m\sigma^2)^{1/2}}\right] e^{-\frac{(y-m\mu)^2}{2m\sigma^2}} \frac{(T\lambda)^m e^{-T\lambda}}{m!}. \tag{4.22}$$

Setting $f''_m(y) = 0$ and dividing off the non-zero terms we have,

$$\begin{aligned} -\frac{1}{m\sigma^2} + \frac{(y-m\mu)^2}{m^2\sigma^4} &= 0 \\ -m\sigma^2 + (y-m\mu)^2 &= 0 \\ y &= m\mu \pm \sqrt{m\sigma^2}. \end{aligned} \tag{4.23}$$

Now,

$$f''_m(y) = \frac{(T\lambda)^m e^{-T\lambda}}{m! m\sigma^2 (2\pi m\sigma^2)^{1/2}} e^{-\frac{(y-m\mu)^2}{2m\sigma^2}} \left[\frac{(y-m\mu)^2}{m\sigma^2} - 1 \right]. \tag{4.24}$$

Since the first two terms are positive, $f''_m(y)$ is negative only on the interval $(m\mu - \sqrt{m\sigma^2}, m\mu + \sqrt{m\sigma^2})$ and $\lim_{y\to\infty} f'_m(y) = \lim_{y\to-\infty} f'_m(y) = 0$, $f_m(y)$ reaches its maximum rate of change at $m\mu - \sqrt{m\sigma^2}$ and its minimum rate of change at $m\mu + \sqrt{m\sigma^2}$. Also, $\left|f'_m(m\mu - \sqrt{m\sigma^2})\right| = \left|f'_m(m\mu + \sqrt{m\sigma^2})\right|$.

Let $y^* = m\mu - \sqrt{m\sigma^2}$, then for fixed $\epsilon > 0$, find a $\delta > 0$ such that $\left|f_m(y^* + \frac{\delta}{2}) - f_m(y^* - \frac{\delta}{2})\right| < \epsilon$, which is possible since $f_m(y)$ is a continuous function of y, then δ is the bound for the absolute continuity of the function $f_m(y)$.

Now, since $0 < \frac{(T\lambda)^m e^{-T\lambda}}{m!} < 1$ for all m, if $f(y)$ is the pdf of $N(\mu, \sigma^2)$ then,

$$f'_m(y^*) = \frac{1}{\sqrt{2\pi m\sigma^2}} e^{-\frac{1}{2}} \frac{(T\lambda)^m e^{-t\lambda}}{m!} < \frac{1}{\sqrt{2\pi m\sigma^2}} e^{-\frac{1}{2}} = f'(\mu - \sqrt{\sigma^2}) \tag{4.25}$$

for all $m = 1, 2, \cdots$. Thus, given $\epsilon > 0$, if δ^* is the bound for the absolute continuity for $f'(y)$ it also is the bound for $f'_m(y)$ for $m = 1, 2, \cdots$, and Part (i) of the theorem is proved.

Part (ii): For $m < 1$, $F(x, m)$ is clearly discrete, so consider $m \geq 1$, since $\Phi(z)$ is the distribution function of a standard normal random variable it is clearly

absolutely continuous, so $F(x,m)$ is the sum of a finite number of absolutely continuous functions and is therefore absolutely continuous, and the lemma is proven. □

Note that, if we let the two parameters, μ and λ, in the distribution to be $\theta = (\mu, \lambda)$, then the C.D.F. $F(x,m)$ is also expressed as $F(x,m,\theta)$ and pdf $f(x,m)$ as $f(x,m,\theta)$ when needed. Furthermore, for any $\theta = (\mu,\lambda)$ and positive number ρ, let $f(y,m,\theta,\rho)$ be the supremum of $f(y,m,\theta')$ with respect to θ' where $|\theta - \theta'| \leq \rho$. For any positive number r, let $\varphi(y,m,r)$ be the supremum of $f(y,m,\theta)$ with respect to θ when $|\theta| > r$. Furthermore, let

$$f^*(y,m,\theta,\rho) = \begin{cases} f(y,m,\theta,\rho) & \text{when} \quad f(y,n,\theta,\rho) > 1 \\ 1 & \text{when} \quad f(y,n,\theta,\rho) \leq 1 \end{cases} \tag{4.26}$$

and

$$\varphi^*(y,m,r) = \begin{cases} \varphi(y,m,r) & \text{when} \quad \varphi(y,m,r) > 1 \\ 1 & \text{when} \quad \varphi(y,m,r) \leq 1 \end{cases} \tag{4.27}$$

We have the following lemma.

LEMMA 4.2. For sufficiently small ρ and for sufficiently large r the expectation of $\log f^*(y,m,\theta,\rho)$ and $\log \varphi^*(y,m,r)$, given θ_0 the true parameter, are finite.

PROOF. First consider the expected value of $\log f^*(y,m,\theta,\rho)$, since $\log f^*(y,0,\theta,\rho) = 0$ for all y,

$$E[\log f^*(y,m,\theta,\rho)] = \sum_{m=1}^{\infty} \int_{-\infty}^{+\infty} \log f^*(y,m,\theta,\rho) dF(y,m,\theta_0), \tag{4.28}$$

where, $dF(y, m, \theta_0) = f(y|\mu_0, m)f(m|\lambda_0)dy$.

So Equation (4.28) can be written as

$$E[\log f^*(y, m, \theta, \rho)] = \sum_{m=1}^{\infty} f(m|\lambda_0) \int_{-\infty}^{+\infty} \log f^*(y, m, \theta, \rho) f(y|\mu_0, m) dy. \qquad (4.29)$$

Now if $f^*(y, m, \theta, \rho) = 1$ for all pairs (y, m), then the result is obtained. So assume there exists an interval of y for which $f^*(y, m, \theta, \rho) \neq 1$ for at least one m. Let $\phi(x)$ be the pdf of a normal variable with mean 0 and variance σ^2, where σ^2 is the same as in the Compound Poisson process. Since, $f(y, m, \theta)$ is the product of a normal pdf, and the pmf of a Poisson random variable which has a value less than one, $f(y, m, \theta) < \phi(0)$ for all θ and $f^*(y, m, \theta, \rho) < \phi(0)$ for all pairs (y,m).

Now since the maximum of a normal function decreases as the variance increases, there exists an $M < \infty$ such that if $m = M$, $f^*(y, m, \theta, \rho) \neq 1$ for some interval of y and if $m > M$, there is no such interval. Now for $1 \leq m \leq M$, let a_m and b_m be the end points of the interval where $f(y, m, \theta, \rho) \neq 1$. By the definition of the pdf, these must be finite. Let $a = \min_{1 \leq m \leq M} a_m$ and $b = \max_{1 \leq m \leq M} b_m$ and consider the integral in Equation (4.29).

$$\begin{aligned} 0 &\leq \int_{-\infty}^{\infty} \log f^*(y, m, \theta, \rho) f(y|\mu_0, m) dy \\ &= \int_{a_n}^{b_n} \log f^*(y, m, \theta, \rho) f(y|\mu_0, m) dy \\ &\leq \int_{a}^{b} \log \phi(0) f(y|\mu_0, m) dy = C < \infty. \qquad (4.30) \end{aligned}$$

Using the result from Equation (4.30) and the fact that for $m > M$, $log f^*(y, m, \theta, \rho) = 0$ the Equation (4.29) continues as follows,

$$E[\log f^*(y,m,\theta,\rho)] \leq C \sum_{m=1}^{M} f(m|\lambda_0) < C < \infty. \tag{4.31}$$

The proof that $E[\log \varphi^*(y,m,r)]$ is finite follows a similar argument, thus the lemma is proven. □

LEMMA 4.3. If $\lim_{i\to\infty} \theta_i = \theta$, then $\lim_{i\to\infty} f(y,m,\theta_i) = f(y,m,\theta)$ for all (y,m).

PROOF. The lemma follows immediately from the fact that $f(y,m)$ is continuous for both μ and λ over the entire parameter space. □

LEMMA 4.4. If θ_1 is a parameter point different from the true parameter point θ_0, then $F(y,m,\theta_1) \neq F(y,m,\theta_0)$ for at least one value of (y,n).

PROOF. Consider the case where $\lambda_1 \neq \lambda_0$, then $F(0,0,\theta_1) = e^{-T\lambda_1} \neq e^{-T\lambda_0} = F(0,0,\theta_0)$.

Only the case where $\lambda_1 = \lambda_0 = \lambda$ and $\mu_1 \neq \mu_0$ need now be considered. Let $F_y(y|\mu,m)$ be the conditional C.D.F. of y given m, then,for $m = 1$

$$F(1,\mu_1,\theta_1) = \frac{1}{e^{T\lambda}} + \frac{T\lambda}{e^{T\lambda}} F_y(\mu_1|\mu = \mu_1, m = 1), \tag{4.32}$$

while

$$F(1,\mu_1,\theta_0) = \frac{1}{e^{T\lambda}} + \frac{T\lambda}{e^{T\lambda}} F_y(\mu_1|\mu = \mu_0, m = 1). \tag{4.33}$$

But they can't be equal since,

$$F_y(\mu_1|\mu = \mu_1, m = 1) = 0.5 \neq F_y(\mu_1|\mu = \mu_0, m = 1). \tag{4.34}$$

Thus the lemma is proved. □

LEMMA 4.5. For the true parameter point θ_0 we have,

$$\sum_{m=1}^{\infty}\int_{-\infty}^{\infty}|\log f(y,m,\theta_0)|dF(y,m,\theta_0)<\infty. \tag{4.35}$$

PROOF. Let θ_0 be the true parameter and consider,

$$E[|\log f(y,m,\theta_0)|]=T\lambda_0 e^{-T\lambda_0}+\sum_{m=1}^{\infty}\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\int_{-\infty}^{\infty}|\log f(y,m,\theta_0)|$$
$$\cdot\frac{1}{\sqrt{2\pi m\sigma^2}}e^{-\frac{(y-m\mu)^2}{2m\sigma^2}}dy. \tag{4.36}$$

Since the first term is finite, it need only be shown the sum is also finite. So begin by simplifying the integral as follows,

$$\int_{-\infty}^{\infty}|\log f(y,m,\theta_0)|\frac{1}{\sqrt{2\pi m\sigma^2}}e^{-\frac{(y-m\mu)^2}{2m\sigma^2}}dy$$

$$= \int_{-\infty}^{\infty}\left|\log\left(\tfrac{1}{\sqrt{2\pi m\sigma^2}}\right)-\tfrac{(y-m\mu_0)^2}{2m\sigma^2}+\log\left(\tfrac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right|\tfrac{1}{\sqrt{2\pi m\sigma^2}}e^{-\frac{(y-m\mu_0)^2}{2m\sigma^2}}dy$$

$$\leq \left|\log\tfrac{1}{\sqrt{2\pi m\sigma^2}}\right|\int_{-\infty}^{\infty}\tfrac{1}{\sqrt{2\pi m\sigma^2}}e^{-\frac{(y-m\mu_0)^2}{2m\sigma^2}}dy+\tfrac{1}{\sqrt{2\pi m\sigma^2}}\int_{-\infty}^{\infty}\tfrac{(y-m\mu_0)^2}{2m\sigma^2}e^{-\frac{(y-m\mu_0)^2}{2m\sigma^2}}dy$$
$$+\left|\log\left(\tfrac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right|\int_{-\infty}^{\infty}\tfrac{1}{\sqrt{2\pi m\sigma^2}}e^{-\frac{(y-m\mu_0)^2}{2m\sigma^2}}dy. \tag{4.37}$$

Let $X=\frac{(y-m\mu_0)}{\sqrt{2m\sigma^2}}$, then $dX=\frac{1}{\sqrt{2m\sigma^2}}dy$, by substituting this into the second term above and integrating the first and third terms Equation (4.37) becomes

$$= \left|\log\frac{1}{\sqrt{2\pi m\sigma^2}}\right|+\frac{2}{\sqrt{\pi}}\int_0^{\infty}X^2e^{-X^2}dX+\left|\log\left(\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right|$$
$$= \left|\log\frac{1}{\sqrt{2\pi m\sigma^2}}\right|+\frac{2}{\sqrt{\pi}}\cdot\frac{\sqrt{\pi}}{4}+\left|\log\left(\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right|$$

$$\begin{aligned} &\leq \left|\log \frac{1}{\sqrt{2\pi\sigma^2}}\right| + \left|\frac{1}{2}\log m\right| + \frac{1}{2} + \left|\log\left(\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right| \\ &= \frac{1}{2} + \left|\frac{1}{2}\log(2\pi\sigma^2)\right| + \frac{1}{2}\log m + \left|\log\left(\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right|. \end{aligned} \tag{4.38}$$

Now substituting Equation (4.38) into the sum in Equation (4.36) it need only be shown that the following sum is finite.

$$\begin{aligned} &\sum_{m=1}^{\infty}\left(\frac{1}{2} + \left|\frac{1}{2}\log(2\pi\sigma^2)\right| + \frac{1}{2}\log m + \left|\log\left(\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right|\right)\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!} \\ = \;& \frac{1}{2} + \left|\frac{1}{2}\log(2\pi\sigma^2)\right| + \frac{1}{2}\sum_{m=1}^{\infty}\frac{(T\lambda_0)^m e^{-T\lambda_0}\log m}{m!} \\ &+ \sum_{m=1}^{\infty}\left|\log\left(\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right|\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!} \end{aligned} \tag{4.39}$$

As the first two terms of the Equation (4.39) are finite, it need only be shown that the last two terms are finite. Using the fact that $\frac{\log m}{m} < 1$, the first of those terms may be simplified as follows,

$$\begin{aligned} \frac{1}{2}\sum_{m=1}^{\infty}\frac{(T\lambda_0)^m e^{-T\lambda_0}\log m}{m!} &< \frac{T\lambda_0}{2}\sum_{m=1}^{\infty}\frac{(T\lambda_0)^{m-1}e^{-T\lambda_0}}{(m-1)!} \\ &= \frac{T\lambda_0}{2}. \end{aligned} \tag{4.40}$$

This leaves only to show that

$$\sum_{m=1}^{\infty}\left|\log\left(\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!}\right)\right|\frac{(T\lambda_0)^m e^{-T\lambda_0}}{m!} \tag{4.41}$$

is finite.

Using the ratio test for power series consider,

$$\lim_{m\to\infty} \frac{\left|\log\left(\frac{(T\lambda_0)^{m+1}e^{-T\lambda_0}}{(m+1)!}\right)\right| \frac{(T\lambda_0)^{m+1}e^{-T\lambda_0}}{(m+1)!}}{\left|\log\left(\frac{(T\lambda_0)^{m}e^{-T\lambda_0}}{m!}\right)\right| \frac{(T\lambda_0)^{m}e^{-T\lambda_0}}{m!}}$$

$$= T\lambda_0 \lim_{m\to\infty} \frac{m!}{(m+1)!} \left|\frac{(m+1)\log(T\lambda_0) - T\lambda_0 - \log(m+1)!}{m\log(T\lambda_0) - T\lambda_0 - \log m!}\right|$$

$$= T\lambda_0 \lim_{m\to\infty} \frac{1}{m+1} \left|\frac{\frac{(m+1)\log(T\lambda_0)}{\log m!} - \frac{T\lambda_0}{\log m!} - \frac{\log(m+1)!}{\log m!}}{\frac{m\log(T\lambda_0)}{\log m!} - \frac{T\lambda_0}{\log m!} - 1}\right| \tag{4.42}$$

Since only the ratio of $\frac{\log(m+1)!}{\log m!}$ and 1 do not go to zero in the absolute value as m goes to infinity. So Equation (4.42),

$$= T\lambda_0 \lim_{m\to\infty} \frac{1}{m+1} \left|\frac{\log(m+1)!}{\log m!}\right|$$

$$= T\lambda_0 \lim_{m\to\infty} \frac{1}{m+1} \left|\frac{\log m!}{\log m!} + \frac{\log(m+1)}{\log m!}\right|$$

$$= 0. \tag{4.43}$$

Thus, the sequence converges for all $T\lambda_0 > 0$ and the lemma is proved. □

LEMMA 4.6. $f(y, m, \theta, \rho)$ is a measurable function of (y, m) for any θ and ρ.

PROOF. Clearly, the domain of (y, m) is measurable and does not change for θ or ρ. Also, as Lemma (4.3) holds for all pairs (y, m) in the domain, $f(y, m, \theta, \rho)$ is measurable for all θ and ρ. □

LEMMA 4.7. Let $\delta(\theta_1, \theta_2)$ be the Euclidian distance in real two space and $\Omega = \{(\mu, \lambda)|(\mu, \lambda) \in (-\infty, \infty) \times (0, \infty)\}$, then the following properties hold,

(i) (Ω, δ) form a metric space,

(ii) If θ_0 is a fixed point in Ω and $\lim_{i\to\infty} \delta(\theta_i, \theta_0) = \infty$, then $\lim_{i\to\infty} f(y, m, \theta_i) = 0$

for any (y, m),

(iii) Any closed and bounded subset of Ω is compact.

PROOF. Part (i), clearly by the properties of real numbers, (Ω, δ) form a metric space.

Part (ii), let θ_0 be a fixed point in Ω and θ_i a sequence of points such that $\lim_{i\to\infty} \delta(\theta_i, \theta_0) = \infty$. Since θ_0 is fixed, this implies $\|\theta_i\| \to \infty$ so either $|\mu| \to \infty$ or $\lambda \to \infty$. But for a fixed point (y, m) if $|\mu| \to \infty$ then $P(y|m) \to 0$ and if $\lambda \to \infty$ then $P(m) \to 0$. Since $f(y, m) = P(y|m) \cdot P(m)$ then $\lim_{i\to\infty} f(y, m, \theta_i) \to 0$ for each (y, m).

Part (iii), by properties of real two space, any closed and bounded subset of Ω is compact. □

Through the rest of this section, let $\theta = (\mu, \lambda)$, $\theta_i = (\mu_i, \lambda_i)$ and $\theta_0 = (\mu_0, \lambda_0)$ the true parameters. E is the expected value with respect to the true θ_0.

LEMMA 4.8. For any $\theta \neq \theta_0$ we have,

$$E \log f(X, M, \theta) < E \log f(X, M, \theta_0), \tag{4.44}$$

where (X, M) are a pair of random variables with a Compound Poisson distribution function $F(x, m, \theta_0)$.

PROOF. It follows from Lemma (4.2) that the expected values in the lemma exist. From Lemma (4.5), we have

$$E|\log f(X, M, \theta_0)| < \infty. \tag{4.45}$$

If $E \log f(X, M, \theta) = -\infty$, the lemma holds. So consider the case where $E \log f(X, M, \theta) > -\infty$. Then

$$E|\log f(X, M, \theta)| < \infty. \tag{4.46}$$

Let $u = \log f(X, M, \theta) - \log f(X, M, \theta_0)$. Then because of Equations (4.46) and (4.47), $E|u| < \infty$. It is well known that for any chance variable u for which $E|u| < \infty$,

$$Eu < \log Ee^u \tag{4.47}$$

Note: This is a generalization of the relation between arithmatic and geometric mean.

Since in our case $Ee^u = E\left(\frac{f(X,M,\theta)}{f(X,M,\theta_0)}\right) \leq 1$ and since u differs from zero on a set of positive probability due to Lemma (4.4), from Equation (4.47) it is obtained that,

$$Eu = E(\log f(X, M, \theta) - \log f(X, M, \theta_0) < 0. \tag{4.48}$$

Thus Lemma (4.8) is proven. □

LEMMA 4.9. $lim_{\rho \to 0} E \log f(X, M, \theta, \rho) = E \log f(X, M, \theta)$.

PROOF. Let,

$$f^{\geq 1}(x, m, \theta, \rho) = \begin{cases} 1 & \text{when} \quad f(x, m, \theta, \rho) < 1 \\ f(x, m, \theta, \rho) & \text{when} \quad f(x, m, \theta, \rho) \geq 1. \end{cases} \tag{4.49}$$

Similarly, let

$$f^{\geq 1}(x, m, \theta) = \begin{cases} 1 & \text{when} \quad f(x, m, \theta) < 1 \\ f(x, m, \theta) & \text{when} \quad f(x, m, \theta) \geq 1. \end{cases} \tag{4.50}$$

It follows from Lemma (4.3) that

$$\lim_{\rho \to 0} \log f^{\geq 1}(x, m, \theta, \rho) = \log f^{\geq 1}(x, m, \theta). \tag{4.51}$$

Since $\Omega = (-\infty, \infty) \times (0, \infty)$, $\{\theta_\rho || \theta_\rho - \theta| \leq \rho\}$ is a compact set whenever $\lambda > \rho$. Since $\log f^*(x, m, \theta, \rho)$ is a non-decreasing function of ρ, it follows from Lemma (4.2) and as $\rho \to 0$, ρ will be less than λ, that

$$\lim_{\rho \to 0} \log f^{\geq 1}(X, M, \theta, \rho) = E \log f^{\geq 1}(X, M, \theta). \tag{4.52}$$

Now, let

$$f^{\leq 1}(x, m, \theta, \rho) = \begin{cases} f(x, m, \theta, \rho) & \text{when} \quad f(x, m, \theta, \rho) \leq 1 \\ 1 & \text{when} \quad f(x, m, \theta, \rho) > 1. \end{cases} \tag{4.53}$$

Similarly, let

$$f^{\leq 1}(x, m, \theta) = \begin{cases} f(x, m, \theta) & \text{when} \quad f(x, m, \theta) \leq 1 \\ 1 & \text{when} \quad f(x, m, \theta) > 1. \end{cases} \tag{4.54}$$

Since $f^{\leq 1}(x, m, \theta, \rho)$ and $f^{\leq 1}(x, m, \theta)$ are both between zero and one and by the definition of $f(x, m, \theta, \rho)$, $f(x, m, \theta, \rho) \geq f(x, mx\theta)$,

$$|\log f^{\leq 1}(x, m, \theta, \rho)| \leq |\log f^{\leq 1}(x, m, \theta) \tag{4.55}$$

and clearly,

$$\lim_{\rho \to 0} \log f^{\leq 1}(x, m, \theta, \rho) = \log f^{\leq 1}(x, m, \theta) \tag{4.56}$$

It follows from Equations (4.55) and (4.56) that,

$$E \log f^{\leq 1}(X, M, \theta, \rho) = E \log f^{\leq 1}(X, M, \theta), \tag{4.57}$$

whether $E \log f^{\leq 1}(X, M, \theta)$ is finite or negative infinity. The Lemma follows directly from Equations (4.52) and (4.57). □

LEMMA 4.10. The equation

$$\lim_{r \to \infty} E \log \varphi(X, M, r) = -\infty \qquad \text{holds.} \tag{4.58}$$

PROOF. Let θ_0 represent the true parameter, then as $r \to \infty$, $\delta(\theta_0, \theta_n) \to \infty$, where $|\theta_n| > r$ and $n - 1 \leq r < n$. Then Lemma (4.7) part (ii), gives

$$\lim_{r \to \infty} \varphi(x, m, r) = 0. \tag{4.59}$$

Which implies that

$$\lim_{r \to \infty} \log \varphi(x, m, r) = -\infty. \tag{4.60}$$

Since according to Lemma (4.2),

$$E \log \varphi^*(X, M, r) < \infty \tag{4.61}$$

and $\log \varphi^*(x, m, r)$ is a decreasing function of r,

$$0 \leq \lim_{r \to \infty} \log \varphi^*(X, M, r) < \infty \tag{4.62}$$

Since, $A(x, m, r) = \log \varphi(x, m, r) - \log \varphi^*(x, m, r)$ is also a decreasing function of r, $E[A(X, M, r)]$ is a decreasing function of r and from Equation (4.60),

$$\lim_{r \to \infty} E[A(X, M, r)] = -\infty. \tag{4.63}$$

Then from Equations (4.62) and (4.63) the Lemma is proven. □

THEOREM 4.11. Let ω be any closed subset of the parameter space Ω which does

not contain the true parameter point θ_0. Then

$$P\left\{\lim_{n\to\infty}\frac{Sup_{\theta\in\Omega}f(X_1,M_1,\theta)f(X_2,M_2,\theta)\cdots f(X_n,M_n,\theta)}{f(X_1,M_1,\theta_0)f(X_2,M_2,\theta_0)\cdots f(X_n,M_n,\theta_0)}=0\right\}=1. \tag{4.64}$$

PROOF. Let r_0 be a positive number chosen such that

$$E\log\varphi(X,M,r_0)<E\log f(X,M,\theta_0). \tag{4.65}$$

The existence of such a positive number follows from Lemma (4.10). Let $\omega_{r_0}=\{\theta|\theta\in\Omega \text{ and } \|\theta|\leq r_0\}$. For each point θ in ω_{r_0} we associate a positive value $|\rho_\theta$ such that

$$E\log f(X,M,\theta,\rho_\theta)<E\log f(X,M,\theta_0). \tag{4.66}$$

The existence of ρ_θ follows from Lemmas (4.8) and (4.9). Since the set ω_{r_0} is compact, their exists a finite number of points $\theta_1,\theta_2,\cdots,\theta_h$ in ω_{r_0} such that $\omega_{r_0}\subseteq S(\theta_1,\rho_{\theta_1})\cup S(\theta_2,\rho_{\theta_2})\cup\cdots\cup S(\theta_h,\rho_{\theta_h})$.

Here $S(\theta,\rho)$ denotes the circle with center at θ and radius ρ. Clearly

$$\begin{aligned}0 \quad\leq\quad & Sup_{\theta\in\omega}f(x_1,m_1,\theta)\cdots f(x_n,m_n,\theta)\leq\sum_{i=1}^{h}f(x_1,m_1,\theta_i,\rho_{\theta_1})\\ & \cdots f(x_n,m_n,\theta_i,\rho_{\theta_1})+\varphi(x_1,m_1,r_0)\cdots\varphi(x_1,m_1,r_0).\end{aligned} \tag{4.67}$$

Hence, Theorem (4.11) is proved if we can show that

$$P\left\{\lim_{n\to\infty}\frac{f(X_1,M_1,\theta_i,\rho_{\theta_i})\cdots f(X_n,M_n,\theta_i,\rho_{\theta_i})}{f(X_1,M_1,\theta_0)\cdots f(X_n,M_n,\theta_0)}=0\right\}=1. \tag{4.68}$$

for $i=1,2,\cdots,h$ and,

$$P\left\{\lim_{n\to\infty}\frac{\varphi(X_1,M_1,r_0)\cdots\varphi(X_n,M_n,r_0)}{f(X_1,M_1,\theta_0)\cdots f(X_n,M_n,\theta_0)}=0\right\}=1. \tag{4.69}$$

The above equations can be written as,

$$P\left\{\lim_{n\to\infty}\sum_{j=1}^{n}\left[\log f(X_j, M_j, \theta_i, \rho_{\theta_i}) - \log f(X_j, M_j, \theta_0)\right] = -\infty\right\} = 1 \tag{4.70}$$

for $i = 1, 2, \cdots, h$ and,

$$P\left\{\lim_{n\to\infty}\sum_{j=1}^{n}\left[\log \varphi(X_j, M_j, r_0) - \log f(X_j, M_j, \theta_0)\right] = -\infty\right\} = 1 \tag{4.71}$$

But these follow directly from Equations (4.65) and (4.66) and the strong law of large numbers. Thus the Theorem (4.11) is proved. □

THEOREM 4.12. Let $\tilde{\theta}_n = g\left((x_1, m_1), (x_2, m_2), \cdots, (x_n, m_n)\right)$ be a function of the observations $(x_1, m_1), (x_2, m_2), \cdots, (x_n, m_n)$ such that

$$\frac{f(x_1, m_1, \tilde{\theta}_n)\cdots f(x_n, m_n, \tilde{\theta}_n)}{f(x_1, m_1, \theta_0)\cdots f(x_n, m_n, \theta_0)} \geq C > 0 \tag{4.72}$$

for all n and for all $(x_1, m_1), (x_2, m_2), \cdots, (x_n, m_n)$. Then

$$P\left(\lim_{n\to\infty}\tilde{\theta}_n = \theta_0\right) = 1. \tag{4.73}$$

PROOF. It is sufficient to prove that for any $\epsilon > 0$ the probability is one that all limit points $\tilde{\theta}$ of the sequence $\{\tilde{\theta}_n\}$ satisfy the inequality $|\tilde{\theta} - \theta_0| \leq \epsilon$.

Now, suppose there exists a limit point $\tilde{\theta}$ of the sequence $\{\tilde{\theta}_n\}$ such that $|\tilde{\theta} - \theta_0| > \epsilon$, then

$$Sup_{|\tilde{\theta}-\theta_0|\geq\epsilon} f(x_1, m_1, \theta)\cdots f(x_n, m_n, \theta) \geq f(x_1, m_1, \tilde{\theta}_n)\cdots f(x_n, m_n, \tilde{\theta}_n) \tag{4.74}$$

for infinitely many n. But then,

$$\frac{Sup_{|\tilde{\theta}-\theta_0|\geq\epsilon} f(x_1, m_1, \theta)\cdots f(x_n, m_n, \theta)}{f(x_1, m_1, \tilde{\theta}_0)\cdots f(x_n, m_n, \tilde{\theta}_0)} \geq \epsilon > 0 \tag{4.75}$$

for infinitely many n. But according to Theorem (4.11), the probability of this

event is zero, thus Theorem (4.12) is proven. □

Since, the MLEs, $\hat{\theta}_n$ satisfies Theorem (4.12) for $C = 1$, thus the consistency of the MLEs are established for the case of known variance of the normal model.

Maximum Likelihood Procedure

Like with all change–point problems, the generalized likelihood ratio test is not appropriate and only the likelihood ratio procedure test, Lehmann (1986) [116], will be used. The test is based on,

$$\Lambda = \frac{L_0(\hat{\mu}, \hat{\lambda})}{\max_{1\leq \jmath \leq \ell-1} L_1(\hat{\mu_1}, \hat{\mu_2}, \hat{\lambda_1}, \hat{\lambda_2})} = \min_{1\leq \jmath \leq \ell-1} \frac{L_0(\hat{\mu}, \hat{\lambda})}{L_1(\hat{\mu_1}, \hat{\mu_2}, \hat{\lambda_1}, \hat{\lambda_2})} \tag{4.76}$$

By canceling the common terms in Equations (4.2) and (4.3) the following equation is established,

$$\begin{aligned}
\Lambda \quad &= \quad \min_{1\leq \jmath \leq \ell-1} \frac{e^{-\frac{1}{2\sigma^2}\sum_{i\in A}\frac{(x_{t_i}-m_i\hat{\mu})^2}{m_i}}}{e^{-\frac{1}{2\sigma^2}\left(\sum_{i\in A_j}\frac{(x_{t_i}-m_i\hat{\mu}_1)^2}{m_i}+\sum_{i\in A_\ell}\frac{(x_{t_i}-m_i\hat{\mu}_2)^2}{m_i}\right)}} \\
&\qquad \cdot \frac{e^{-\hat{\lambda}\sum_{i=1}^{\ell} T_i}\hat{\lambda}^{\sum_{i=1}^{\ell} m_i}}{e^{-\left(\hat{\lambda_1}\sum_{i=1}^{j} T_i+\hat{\lambda_2}\sum_{i=j+1}^{\ell} T_i\right)}\hat{\lambda_1}^{\sum_{i=1}^{j} m_i}\cdot \hat{\lambda_2}^{\sum_{i=j+1}^{\ell} m_i}} \\
= \quad &\min_{1\leq \jmath \leq \ell-1} \left[\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i\in A}\frac{(x_{t_i}-m_i\hat{\mu})^2}{m_i} - \sum_{i\in A_j}\frac{(x_{t_i}-m_i\hat{\mu}_1)^2}{m_i}\right.\right.\right. \\
&\qquad \left.\left.\left. - \sum_{i\in A_\ell}\frac{(x_{t_i}-m_i\hat{\mu}_2)^2}{m_i}\right)\right\}\right. \\
&\qquad \left.\cdot \exp\left\{-\left(\hat{\lambda}\sum_{i=1}^{\ell} T_i - \hat{\lambda_1}\sum_{i=1}^{j} T_i - \hat{\lambda_2}\sum_{i=j+1}^{\ell} T_i\right)\right\}\right.
\end{aligned}$$

$$\cdot \frac{\hat{\lambda}^{\sum_{i=1}^{\ell} m_i}}{\hat{\lambda_1}^{\sum_{i=1}^{j} m_i} \cdot \hat{\lambda_2}^{\sum_{i=j+1}^{\ell} m_i}} \Bigg]. \tag{4.77}$$

Since,

$$\hat{\lambda} \sum_{i=1}^{\ell} T_i = \sum_{i=1}^{\ell} m_i, \; \hat{\lambda_1} \sum_{i=1}^{j} T_i = \sum_{i=1}^{j} m_i, \text{ and } \hat{\lambda_2} \sum_{i=j+1}^{\ell} T_i = \sum_{i=j+1}^{\ell} m_i,$$

$$\text{so } \exp\left\{ -\left(\hat{\lambda} \sum_{i=1}^{\ell} T_i - \hat{\lambda_1} \sum_{i=1}^{j} T_i - \hat{\lambda_2} \sum_{i=j+1}^{\ell} T_i \right) \right\} = e^0 = 1, \tag{4.78}$$

and Λ is given by the expression

$$\begin{aligned} \Lambda \quad = \quad & \min_{1 \le \mathrm{J} \le \ell - 1} \Bigg[\exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{i \in A} \frac{(x_{t_i} - m_i \hat{\mu})^2}{m_i} - \sum_{i \in A_j} \frac{(x_{t_i} - m_i \hat{\mu}_1)^2}{m_i} \right. \right. \\ & \left. \left. - \sum_{i \in A_\ell} \frac{(x_{t_i} - m_i \hat{\mu}_2)^2}{m_i} \right) \right\} \cdot \frac{\hat{\lambda}^{\sum_{i=1}^{\ell} m_i}}{\hat{\lambda_1}^{\sum_{i=1}^{j} m_i} \cdot \hat{\lambda_2}^{\sum_{i=j+1}^{\ell} m_i}} \Bigg]. \end{aligned} \tag{4.79}$$

For purposes of a search criteria for locating the position of the change point, the $-2\log(\Lambda)$ was used.

$$\begin{aligned} & -2\log(\Lambda) \\ = & \max_{1 \le \mathrm{J} \le \ell - 1} \left\{ \frac{1}{\sigma^2} \left(\sum_{i \in A} \frac{(x_{t_i} - m_i \hat{\mu})^2}{m_i} - \sum_{i \in A_j} \frac{(x_{t_i} - m_i \hat{\mu}_1)^2}{m_i} - \sum_{i \in A_\ell} \frac{(x_{t_i} - m_i \hat{\mu}_2)^2}{m_i} \right) \right. \\ & \left. -2\log(\hat{\lambda}) \sum_{i=1}^{\ell} m_i + 2\log(\hat{\lambda_1}) \sum_{i=1}^{j} m_i + 2\log(\hat{\lambda_2}) \sum_{i=j+1}^{\ell} m_i \right\}. \end{aligned} \tag{4.80}$$

Now, let S be the sums from the normal distribution, then

$$\begin{aligned} S \quad = \quad & \sum_{i \in A} \frac{(x_{t_i} - m_i \hat{\mu})^2}{m_i} - \sum_{i \in A_j} \frac{(x_{t_i} - m_i \hat{\mu}_1)^2}{m_i} - \sum_{i \in A_\ell} \frac{(x_{t_i} - m_i \hat{\mu}_2)^2}{m_i} \\ = \quad & \sum_{i \in A} \left(\frac{x_{t_i}^2}{m_i} - 2x_{t_i} \hat{\mu} + m_i \hat{\mu}^2 \right) - \sum_{i \in A_j} \left(\frac{x_{t_i}^2}{m_i} - 2x_{t_i} \hat{\mu}_1 + m_i \hat{\mu}_1^2 \right) \end{aligned}$$

$$- \sum_{i \in A_\ell} \left(\frac{x_{t_i}^2}{m_i} - 2x_{t_i}\hat{\mu}_2 + m_i\hat{\mu}_2^2 \right). \tag{4.81}$$

Now since $A = A_j \bigcup A_\ell$ the first terms of each term cancel each other out, leaving

$$\begin{aligned} S &= \sum_{i \in A} \left(-2x_{t_i}\hat{\mu} + m_i\hat{\mu}^2\right) - \sum_{i \in A_j} \left(-2x_{t_i}\hat{\mu}_1 + m_i\hat{\mu}_1^2\right) - \sum_{i \in A_\ell} \left(-2x_{t_i}\hat{\mu}_2 + m_i\hat{\mu}_2^2\right) \\ &= -2\hat{\mu}\sum_{i \in A} x_{t_i} + \hat{\mu}^2 \sum_{i \in A} m_i + 2\hat{\mu}_1 \sum_{i \in A_j} x_{t_i} - \hat{\mu}_1^2 \sum_{i \in A_j} m_i + 2\hat{\mu}_2 \sum_{i \in A_\ell} x_{t_i} - \hat{\mu}_2^2 \sum_{i \in A_\ell} m_i \\ &= -\hat{\mu}^2 \sum_{i \in A} m_i + \hat{\mu}_1^2 \sum_{i \in A_j} m_i + \hat{\mu}_2^2 \sum_{i \in A_\ell} m_i. \end{aligned} \tag{4.82}$$

Substituting Equation (4.82) into Equation (4.80), the test statistic for the location of the change–point is,

$$\begin{aligned} -2\log(\Lambda) = \max_{1 \le \mathrm{J} \le \ell - 1} \Bigg\{ & \frac{1}{\sigma^2} \left(-\hat{\mu}^2 \sum_{i \in A} m_i + \hat{\mu}_1^2 \sum_{i \in A_j} m_i + \hat{\mu}_2^2 \sum_{i \in A_\ell} m_i. \right) \\ & -2\log(\hat{\lambda}) \sum_{i=1}^{\ell} m_i + 2\log(\hat{\lambda_1}) \sum_{i=1}^{j} m_i + 2\log(\hat{\lambda_2}) \sum_{i=j+1}^{\ell} m_i \Bigg\}. \end{aligned} \tag{4.83}$$

This was initially done with the hope that a p-value for the test could be derived, but to date it remains elusive. The $-2\log(\Lambda)$ is also important in the information approach developed in the next section.

Information Approach

Information Criterion have been used for model selection problems for many years. Its primary purpose is to aid in the selection of competing models when the models are of different complexities. The complexity is typically measured be the number of parameters being estimated in the model. Let θ be the parameters in the

model to be estimated, and the idea is to penalize a more complex model, meaning that to add more parameters to the model, it must produce an improvement in the negative two log maximum likelihood function, denoted $-2\log L(\hat{\theta})$, that exceed some threshold. As an example the Akaike information criterion, Akailke (1973)[1], and Swartz information criterion, Swartz (1978)[187] for non change point models are defined as

$$\begin{aligned} AIC &= -2\log L(\hat{\theta}) + 2dim(\hat{\theta}) \\ SIC &= -2\log L(\hat{\theta}) + dim(\hat{\theta})\log(n), \end{aligned} \tag{4.84}$$

where $\hat{\theta}$ is the maximum point of $\log L(\theta)$. The model with the lowest AIC or BIC is then the model selected.

While both these methods have been applied to the change point problem successfully, in the case where the number of change points is fixed, the role the penalty plays is in reducing the false positive rate by creating a higher threshold for improvement than a method based solely on the likelihood function would require. This is evident in the fact that for a fixed number of change points n, the competing models for the locations all have the same number of parameters. Since the estimated position is based on the location with the minimum, AIC or SIC, that is based solely on the behavior of the likelihood function. Meaning, if the location is estimated by finding the position which maximizes the likelihood function, that same location will minimize both the AIC and SIC.

In Chen et al. (2006)[22], it is suggested that the complexity of the model for change point should include a measure of the proximity of the change points to each

other or the ends of the sequence. They propose the following Modified information criterion MIC:

$$MIC(j) = -2\log(\hat{\theta}_1, \hat{\theta}_2, j) + \left[2dim(\theta) + \left(\frac{2k}{n} - 1\right)^2\right]\log(n). \tag{4.85}$$

Under their model, the additional penalty is eliminated for a change point at $k = \frac{n}{2}$ and increases as the location moves to the ends of the sequence.

In the case of the compound Poisson model, the rate at which the events occur is of keen interest. This is especially the case in analyzing aCGH data as indicated by Levin et al. (2005)[117]. For that reason, a new modified information criterion for compound Poisson data, CPIC, is proposed. Under the proposed criterion, the penalty on each location is reduced for positions which create changes in λ. The proposed complexity is given as:

$$\left(2dim(\theta) + \frac{1}{1 + \sum_{i=1}^{2}\left|\hat{\lambda} - \hat{\lambda}_i\right|}\right)\log(n). \tag{4.86}$$

Under H_0, the hypothesis of no change, the $CPIC(n)$ is calculated by:

$$\begin{aligned} CPIC(n) &= -2\log L_0(\hat{\mu}, \hat{\lambda}) + 2\log n, \\ &= M_A \log(2\pi\sigma^2) + \sum_{i\in A}\log(m_i) + \frac{1}{\sigma^2}\sum_{i\in A}\frac{(x_{t_i} - m_i\hat{\mu})^2}{m_i} \\ &\quad -2\sum_{i=1}^{\ell} m_i \log(T_i) + 2\sum_{i=1}^{\ell}\log(m_i!) - 2\log(\hat{\lambda})\sum_{i=1}^{\ell} m_i \\ &\quad +2\hat{\lambda}\sum_{i=1}^{\ell} T_i + 2\log(n), \end{aligned} \tag{4.87}$$

where $\hat{\mu}$ and $\hat{\lambda}$ are the mle's of μ and λ respectively and found to be:

$$\hat{\mu} = \frac{\sum_{i=1}^{\ell} x_{t_i}}{\sum_{i=1}^{\ell} m_i}, \text{ and } \hat{\lambda} = \frac{\sum_{i=1}^{\ell} m_i}{\sum_{i=1}^{\ell} T_i}. \tag{4.88}$$

Under H_1, the hypothesis of one change point, the $CPIC$, is denoted as $CPIC(j)$ for $j = 1, 2, \cdots, \ell - 1$ is calculated by:

$$CPIC(j) = -2\log L_1(\hat{\mu}_1, \hat{\mu}_2, \hat{\lambda}_1, \hat{\lambda}_2) + \left(2dim(\theta) + \frac{1}{1 + \sum_{i=1}^{2} \left|\hat{\lambda} - \hat{\lambda}_i\right|}\right) \log(n),$$

so,

$$\begin{aligned} CPIC(j) = {} & M_A \log(2\pi\sigma^2) + \sum_{i \in A} \log(m_i) + \frac{1}{\sigma^2} \sum_{i \in A_j} \frac{(x_{t_i} - m_i\hat{\mu}_1)^2}{m_i} \\ & -2\sum_{i=1}^{\ell} m_i \log(T_i) + 2\sum_{i=1}^{\ell} \log(m_i!) - 2\log(\hat{\lambda}_1) \sum_{i=1}^{j} m_i \\ & +2\hat{\lambda}_1 \sum_{i=1}^{j} T_i + \frac{1}{\sigma^2} \sum_{i \in A_\ell} \frac{(x_{t_i} - m_i\hat{\mu}_2)^2}{m_i} - 2\log(\hat{\lambda}_2) \sum_{i=j+1}^{\ell} m_i \\ & +2\hat{\lambda}_2 \sum_{i=j+1}^{\ell} T_i + \left(2dim(\theta) + \frac{1}{1 + \sum_{i=1}^{2} \left|\hat{\lambda} - \hat{\lambda}_i\right|}\right) \log(n). \end{aligned} \tag{4.89}$$

where $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the mle's of μ_1, μ_2, λ_2, and λ_2 respectively and found to be:

$$\begin{aligned} \hat{\mu_1} &= \frac{\sum_{i=1}^{j} x_{t_i}}{\sum_{i=1}^{j} m_i}, \ \hat{\mu_2} = \frac{\sum_{i=j+1}^{\ell} x_{t_i}}{\sum_{i=j+1}^{\ell} m_i}, \\ \hat{\lambda_1} &= \frac{\sum_{i=1}^{j} m_i}{\sum_{i=1}^{j} T_i}, \ \hat{\lambda_2} = \frac{\sum_{i=j+1}^{\ell} m_i}{\sum_{i=j+1}^{\ell} T_i}. \end{aligned} \tag{4.90}$$

The test for change is then to reject H_0 if

$$CPIC(n) > \min_{1 \le j \le \ell - 1} CPIC(j), \tag{4.91}$$

and fail to reject H_0 if

$$CPIC(n) \leq \min_{1 \leq j \leq \ell - 1} CPIC(j). \tag{4.92}$$

If H_0 is rejected, the estimated change point location is

$$CPIC(\hat{j}) = \min_{1 \leq j \leq \ell - 1} CPIC(j). \tag{4.93}$$

Simulation Study

To demonstrate the usefulness of the estimation methods for locating changes in the Compound Poisson model, an extensive simulation study was performed for both the Likelihood Method and the CPIC. The simulation was run for two types of data. The first would simulate a general case of the Compound Poisson process that is a normal random variable paired with a Poisson random variable. To construct this type of data, the Poisson random numbers are generated first and then the normal random variables are generated using the conditional distribution, based on the Poisson variable. This would be indicative of data collected at uniform periods; in the simulation it was assumed to be a length of one. The second type of data models the aCGH data introduced in previous chapters. In this case, both the measurement of the normal variable and the distance between occurrences is collected. Under this situation, the Poisson variable is equal to one for each interval, so all the normal variables under the null hypothesis are identically distributed; while the distances between occurrences are distributed exponentially. So the second type of data is a normal, exponential pair.

Four sets of simulations were one for each combination of data type and method used. Two sets of simulation data was constructed, one for each type of data. Each set contained a change in location, in one of three locations at the n/4th, n/2th, or

3n/4th position. The initial value of the parameters for all sets was $(\mu_0, \lambda_0) = (0, 1)$ and the standard deviation of the normal model equal to one. Three levels of change were considered, $(\mu_2, \lambda_2) = (2, 2)$, (3, 3), and (3.5, 4), and sample sizes used were n = 40, 75, 100, 125, and 200. For each combination of data type, location of change, level of change, and sample size, 1000 data sets were generated. For each case, the frequency of selecting the correct location, the average location of change identified and the MSE of the location are given. A summary of the results can be found in Tables 1 to 6, located at the end of this section.

While no theoretical p-value has been derived for either test, trials performed during this simulation with the Chi Squared distribution with 2 degrees of freedom showed it was too sensitive for the test, when run on sets with no change, it gave a 50% to 80% false positive rate. For the Likelihood method, an empirical p-value was approximated using a weighted mixed model where the square root of the sum of the squared deviations for the normal variables were assigned a p-value using a Gumbel distribution and the exponential or Poisson portion was given a p-value using a Chi Squared distribution with 1 degree of freedom. The p-value was calculated by using 0.5 times their sum. This method reduced the false positive rate to below 1% and gave a reasonable p-value for the tests with change. For the CPIC, a p-value was approximated using a Gumbel distribution of the form used in Chen and Gupta (1997)[24]. The Information Criterion did have the advantage that even without assigning a p-value it had no false positives in the null data sets.

The results of the simulation study showed that both the Likelihood Method and Information Criterion worked equally well. For the normal/exponential data,

each method found between 63.5% and 67.6% of small changes, between 86.3% and 89.3% of medium changes, and between 92.0% and 94.6% of large changes. For the normal/Poisson data, each method found between 53.1% and 59.4% of small changes, between 80.4% and 85.4% of medium changes, and between 89.7% and 93.8% of large changes. Some of the decreased performance of the methods with the normal/Poisson data is with a choice of lambda equal to one there are many cases of $n_i = 0$ and they take away from the points used to estimate μ.

Both tests worked equally well with all three change positions and, as would be expected, increased in efficiency as the level of change increased. What was somewhat surprising is that neither test increased greatly in their efficiency at locating changes, for a fixed level of change, as the sample size increased. For instance, in the normal/exponential results for (μ_2, λ_2) = (3,3) with the change at the n/2th position, using the likelihood method, f=0.886, when n=40 and f=.892, when n=200.

Looking at the average location identified they are close to the actual location with the exception of the normal/Poisson model with n=40, (μ_2, λ_2) = (2,2), and the loci at 10, where it is off by 0.859 and 0.846 for the Likelihood method and the CPIC respectively. For both data types the average gets closer to the actual value as the level of change increases but not necessarily with the number of observations. For all levels and position of change and sample size, the models produce averages that are closer to the actual values for the normal/exponential data. The results for the MSE are much the same as with the averages. The better performance with the normal/exponential data can clearly be seen in the MSEs. The one thing that is notable about the MSEs as compared to the other results which have been discussed

is in the improved values of the MSEs as the sample size increases for low level of changes. This is also evident for the medium level of change but not apparent in the large level of change.

Table 1. Normal/Poisson Data with $(\mu_2, \lambda_2) = (2, 2)$

		LRT Method			CPIC Method		
		Change Position			Change Position		
Sample Size		n/4	n/2	3n/4	n/4	n/2	3n/4
	f	0.531	0.567	0.545	0.532	0.570	0.547
n=40	Mean	10.859	20.492	29.985	10.846	20.528	30.082
	MSE	11.571	5.274	6.061	10.356	5.284	5.796
	f	0.555	0.557	0.556	0.556	0.555	0.550
n=75	Mean	19.356	38.344	56.171	19.359	38.380	56.203
	MSE	4.654	4.800	3.917	4.641	4.724	4.139
	f	0.567	0.557	0.534	0.569	0.557	0.545
n=100	Mean	25.284	50.315	75.181	25.294	50.320	75.208
	MSE	3.762	3.559	3.927	3.718	3.566	3.780
	f	0.552	0.582	0.569	0.554	0.583	0.569
n=125	Mean	31.377	63.293	94.284	31.362	63.291	94.293
	MSE	3.801	4.087	3.156	3.844	4.091	3.141
	f	0.562	0.594	0.575	0.561	0.594	0.576
n=200	Mean	50.362	100.256	150.214	50.369	100.263	150.229
	MSE	3.800	3.234	2.996	3.721	3.199	3.075

Table 2. Normal/Poisson Data with $(\mu_2, \lambda_2) = (3, 3)$

		LRT Method			CPIC Method		
		Change Position			Change Position		
Sample Size		n/4	n/2	3n/4	n/4	n/2	3n/4
	f	0.806	0.843	0.820	0.806	0.844	0.821
n=40	Mean	10.212	20.122	30.052	10.212	20.124	30.064
	MSE	1.230	0.606	0.494	1.230	0.620	0.474
	f	0.829	0.834	0.815	0.830	0.834	0.816
n=75	Mean	19.155	38.126	56.115	19.154	38.126	56.119
	MSE	0.419	0.322	0.479	0.418	0.322	0.463
	f	0.805	0.818	0.844	0.804	0.818	0.845
n=100	Mean	25.125	50.105	75.100	25.124	50.105	75.107
	MSE	0.489	0.467	0.430	0.490	0.467	0.413
	f	0.830	0.816	0.854	0.830	0.816	0.854
n=125	Mean	31.088	63.086	94.031	31.088	63.086	94.031
	MSE	0.382	0.500	0.341	0.382	0.500	0.341
	f	0.843	0.840	0.838	0.843	0.840	0.840
n=200	Mean	50.082	100.105	150.119	50.082	100.105	150.122
	MSE	0.302	0.329	0.301	0.302	0.329	0.304

Table 3. Normal/Poisson Data with $(\mu_2, \lambda_2) = (3.5, 4)$

		LRT Method			CPIC Method		
		Change Position			Change Position		
Sample Size		n/4	n/2	3n/4	n/4	n/2	3n/4
	f	0.897	0.917	0.934	0.897	0.919	0.934
n=40	Mean	10.097	20.040	30.022	10.097	20.036	30.022
	MSE	0.271	0.164	0.096	0.271	0.154	0.096
	f	0.938	0.928	0.911	0.938	0.928	0.912
n=75	Mean	19.046	38.035	56.038	19.046	38.035	56.039
	MSE	0.164	0.113	0.134	0.164	0.113	0.133
	f	0.935	0.923	0.926	0.935	0.923	0.926
n=100	Mean	25.026	50.031	75.040	25.026	50.031	75.042
	MSE	0.140	0.133	0.104	0.140	0.133	0.104
	f	0.914	0.921	0.925	0.914	0.921	0.924
n=125	Mean	31.033	63.044	94.040	31.033	63.044	94.041
	MSE	0.133	0.106	0.116	0.133	0.106	0.117
	f	0.920	0.916	0.928	0.920	0.916	0.928
n=200	Mean	50.044	100.040	150.046	50.044	100.040	150.046
	MSE	0.140	0.128	0.138	0.140	0.128	0.138

Table 4. Normal/Exponential Data with $(\mu_2, \lambda_2) = (2, 2)$

		LRT Method			CPIC Method		
		Change Position			Change Position		
Sample Size		n/4	n/2	3n/4	n/4	n/2	3n/4
	f	0.656	0.662	0.639	0.663	0.662	0.635
n=40	Mean	10.147	20.056	29.857	10.135	20.060	29.899
	MSE	1.879	2.046	3.573	1.769	2.094	3.443
	f	0.664	0.674	0.661	0.662	0.675	0.659
n=75	Mean	19.106	37.979	55.990	19.131	37.983	56.004
	MSE	1.536	1.419	1.544	2.581	1.305	1.582
	f	0.650	0.664	0.641	0.652	0.663	0.643
n=100	Mean	25.092	49.988	74.995	25.106	49.997	75.032
	MSE	1.938	1.118	1.931	1.844	1.207	1.596
	f	0.646	0.664	0.639	0.646	0.664	0.640
n=125	Mean	30.972	62.995	93.947	30.965	62.992	93.952
	MSE	1.478	1.245	1.413	1.483	1.242	1.406
	f	0.652	0.663	0.675	0.649	0.661	0.676
n=200	Mean	50.026	99.985	149.985	50.031	99.984	149.983
	MSE	1.618	1.463	1.179	1.623	1.468	1.175

Table 5. Normal/Exponential Data with $(\mu_2, \lambda_2) = (3, 3)$

		LRT Method			CPIC Method		
		Change Position			Change Position		
Sample Size		n/4	n/2	3n/4	n/4	n/2	3n/4
	f	0.866	0.855	0.881	0.863	0.884	0.874
n=40	Mean	10.003	20.020	29.986	10.006	20.018	29.986
	MSE	0.195	0.228	0.264	0.198	0.246	0.282
	f	0.891	0.876	0.892	0.893	0.874	0.890
n=75	Mean	19.016	38.009	55.980	19.011	38.005	55.982
	MSE	0.194	0.213	0.166	0.181	0.221	0.168
	f	0.884	0.865	0.884	0.884	0.864	0.885
n=100	Mean	25.010	49.996	75.000	25.010	49.994	75.000
	MSE	0.152	0.238	0.186	0.152	0.242	0.188
	f	0.864	0.878	0.887	0.863	0.882	0.887
n=125	Mean	30.978	62.969	93.993	30.979	62.974	93.995
	MSE	0.254	0.183	0.193	0.255	0.176	0.193
	f	0.888	0.892	0.889	0.888	0.892	0.889
n=200	Mean	49.990	99.989	149.970	49.990	99.989	149.970
	MSE	0.160	0.193	0.172	0.160	0.193	0.172

Table 6. Normal/Exponential Data with $(\mu_2, \lambda_2) = (3.5, 4)$

		LRT Method			CPIC Method		
		Change Position			Change Position		
Sample Size		n/4	n/2	3n/4	n/4	n/2	3n/4
	f	0.931	0.920	0.942	0.931	0.920	0.941
n=40	Mean	10.007	19.995	30.015	10.007	19.995	30.016
	MSE	0.101	0.115	0.105	0.101	0.115	0.106
	f	0.939	0.932	0.939	0.927	0.933	0.940
n=75	Mean	18.990	38.002	56.001	18.990	38.004	56.004
	MSE	0.082	0.088	0.081	0.082	0.084	0.080
	f	0.927	0.926	0.940	0.927	0.926	0.939
n=100	Mean	25.092	49.988	74.995	25.106	49.997	75.032
	MSE	0.091	0.105	0.084	0.091	0.105	0.085
	f	0.946	0.923	0.938	0.946	0.923	0.937
n=125	Mean	30.994	63.011	93.987	30.994	63.011	93.988
	MSE	0.084	0.103	0.071	0.084	0.103	0.072
	f	0.938	0.929	0.932	0.938	0.929	0.932
n=200	Mean	49.971	100.002	149.985	49.971	100.002	149.987
	MSE	0.103	0.106	0.089	0.103	0.106	0.089

Application to aCGH Data

Both the likelihood and information methods developed in this section were applied to aCGH data of nine Fibroblast Cell lines with known changes in at least one chromosome. The data sets were part of a set of 15 Fibroblast Cell lines that have been extensively experimented on and the position of the copy number alterations were verified by karyotyping in Snijders et al. (2001)[178]. The results are available at http://nature.com/ng/journal/v29/n3/full/ng754.html. Because of the extensive verification done on these data sets, they are used as a benchmark for studying new methods as in Chen et al. (2010)[30].

The data from each cell line was prepared by first removing data points with missing $\log_2(T_i/G_i)$ values or missing gene position. The data sets included the genome order along the DNA strand, as a marker of the genes location, the chromosome number of each gene, it's $\log_2(T_i/G_i)$, the distances between the gene and the previous gene on the chromosome were calculated using the gene's position on the chromosome, and the number of genes in each interval. The intervals were set up with one gene per interval except for locations with multiple genes, in these cases the $\log_2(T_i/G_i)$'s for the points were combined and the number of genes combined was recorded. Each method was then used to process the data using R. The cell lines were, searched one chromosome at time for possible copy number variations, and a possible location of change for each chromosome was reported. For the likelihood method, a report was generated with a location of possible change, using the genome order data, for each chromosome along with a p-value based on the mixed model, a Chi squared with two degrees of freedom, and a Gumbel distribution. For the information

criterion, a report including the possible location of the change for each chromosome if one was indicated by the CPIC. If no change was indicated on the chromosome then No Change was output. In Figure 1 one can see the results of Chromosome 3 of cell line GM03563, which contains one change point. In the figures, the $Log_2(T_i, G_i)$ are graphed against the position along the gene, with the change point located by both methods denoted with a red circle.

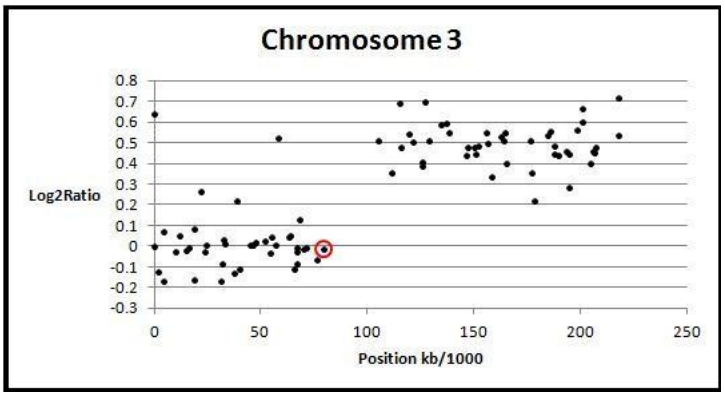

Figure 1. GM03563 Chromosome 3

The results for known copy number locations are given in Table 7. As can be seen, both methods are efficient in finding the location of change, even in the presence of multiple change points. While both methods worked equally as well in locating changes, the CPIC did find the location near the beginning of Chromosome 16 cell line GM04435 as noted by the red diamond in Figure 2, which was missed by the likelihood method. Figure 3 shows the results from cell line GM01524 Chromosome 6, a chromosome with two change points, where one of the two locations were found

by both methods. Figure 4 on the other hand shows were both methods failed to find a change on Chromosome 8 of cell line GM03134. It should be noted, that the location found was adjacent to an outlier, which may have confounded the methods.

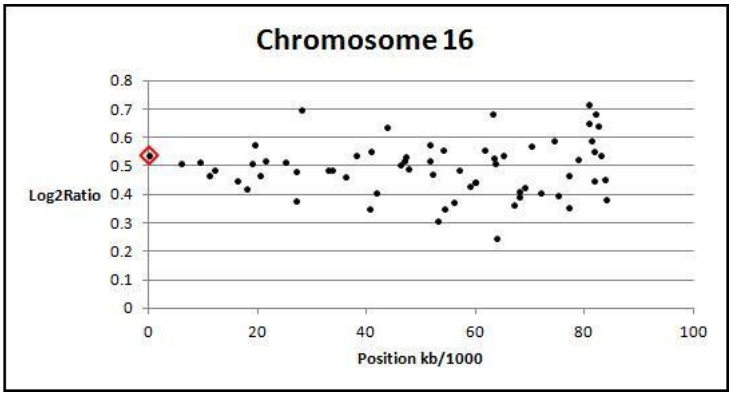

Figure 2. GM04435 Chromosome 16

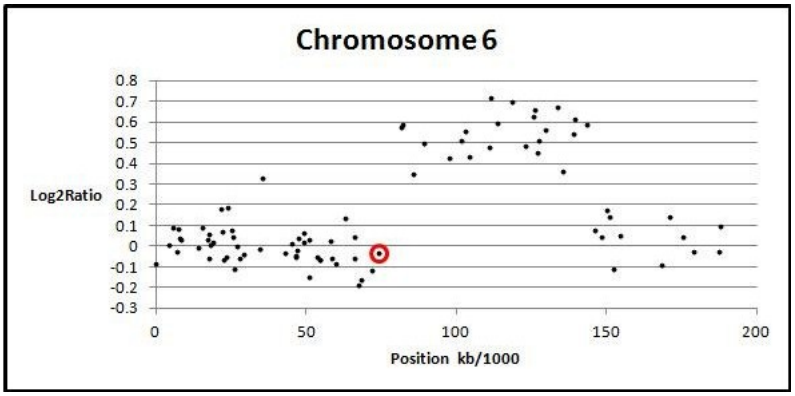

Figure 3. GM01524 Chromosome 6

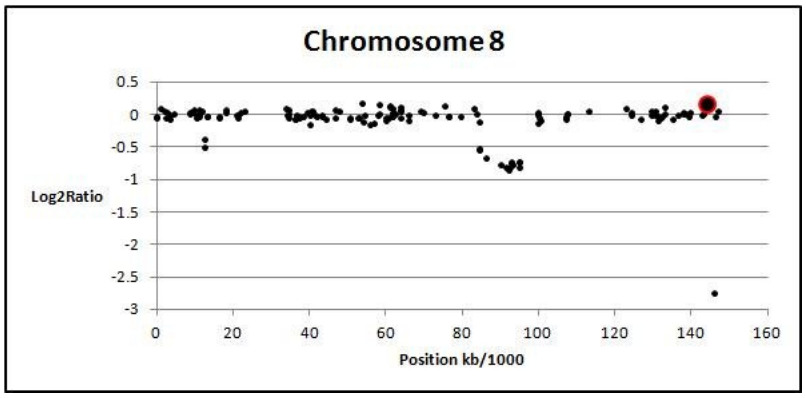

Figure 4. GM013134 Chromosome 8

Of major concern with all applications is the false positive rate. Using the mixed model to calculate a p-value for the likelihood function had a low false positive rate of about 3% for a 10% level of significance and no false positives for the 5% level of significance. However, since the mixed model measures changes in both the μ and λ if the change only occurs in one of the parameters, it also failed to find the correct location a significant change over half the time, even at the 10% level of significance. The information criterion on the other hand indicated change on all known change locations but had over a 50% false positive rate. Clearly the false positive rate on the information criterion could be reduced if a distribution for its test statistic could be found.

Table 7. Outcome for Known Locations of Change $\alpha = 0.1$

			Method			
			Likelihood		CPIC	
Cell Line	Gene	Chrom.	Found	Sign.	Found	Sign.
GM01524	CTD-2009co6	6	Yes	Yes	Yes	Yes
GM01524	RP11-139o22	6	No		No	
GM01535	PR11-88j19	5	Yes	No	Yes	Yes
GM01535	RP11-81G12	12	Yes	No	Yes	Yes
GM01750	RP11-33o15	9	Yes	No	Yes	Yes
GM01750	RP11-125a05	14	Yes	Yes	Yes	Yes
GM03134	RP11-117N14	8	No		No	
GM03134	RP11-102K07	8	No		No	
GM03563	RP11-146e16	3	Yes	Yes	Yes	Yes
GM03563	RP11-28n06	9	Yes	No	Yes	Yes
GM04435	GS1-31J3	5	No		No	
GM04435	CDT-2371a5	16	No		Yes	Yes
GM04435	PAC 191p24	16	No		No	
GM05296	RP11-237j07	10	Yes	No	Yes	Yes
GM05296	RP11-46g13	10	No		No	
GM05296	RP11-127K23	11	No		No	
GM05296	RP11-18B09	11	Yes	Yes	Yes	Yes
GM07081	RP110251—15	7	Yes	Yes	Yes	Yes
GM13330	RP11-234m03	1	Yes	No	Yes	Yes
GM13330	RP11-272o03	4	Yes	Yes	Yes	Yes

CHAPTER 5

CHANGE POINTS IN COMPOUND POISSON PROCESSES WITH VARIANCE UNKNOWN

The Problem

In chapter 4, the case of an abrupt change in mean and lambda which occurred after time j was considered. In this chapter a slightly different problem will be investigated. First, is the possibility of a change in the variance of the normal model σ^2. Second, in the previous problem the interval j was the last interval before the change occurred, in this chapter the change point, interval j, may be distributed differently from all the other intervals in the sequence. This is an important feature of the method developed because, as is noted in Levin et al. (2005)[117], there may be a short abrupt change before the new state is achieved. For instance, in the stock market, there may be an event which causes an initial shock in the price of a stock and after that, the stock may settle on a new growth path. In the case of gene expression data, as discussed in Levin et al. (2005)[117], changes in the expression level of a particular DNA sequence can effect the expression level of neighboring gene. Thus, if a region of the DNA sequence has a changed expression level, it may be that the end of effected sequence may only show a slight change. The model for this case may be expressed as:

Assume that a sequence of observations $y_1, y_2, \ldots, y_n$, are distributed $y_i \sim N(\mu_i, \sigma_i^2)$, and occur during a time or distance T. Let T be subdivided into ℓ intervals

with $t_1, t_2, \ldots, t_\ell$ be the lengths of each interval and N_{t_i} the number of occurrences of y in segment i. Suppose that $N_{t_j} \sim Poisson(\lambda_i t_i)$, and let $X_{t_j} = \sum_{i=1}^{N_{t_j}} y_i$, then $X_{t_1}, X_{t_2}, \ldots, X_{t_\ell}$ form a Compound Poisson Process, and the model for this change point problems can be expressed as:

$$X_{t_i} \sim N(N_{t_i}\mu_1, N_{t_i}\sigma_1^2), \text{ given } N_{t_i}, \text{ for } (i = 1, 2, \cdots, j-1),$$
$$X_{t_j} \sim N(N_{t_j}\delta, N_{t_j}\sigma^2), \text{ given } N_{t_j}$$
$$X_{t_i} \sim N(N_{t_i}\mu_2, N_{t_i}\sigma_2^2), \text{ given } N_{t_i}, \text{ for } (i = j+1, j+2, \cdot, \ell),$$
$$N_{t_i} \sim Poisson(\lambda_1 t_i) \text{ for } (i = 1, 2, \cdots, j-1),$$
$$N_{t_i} \sim Poisson(\lambda t_i) \text{ for } (i = j),$$
$$N_{t_i} \sim Poisson(\lambda_2 t_i) \text{ for } (i = j+1, j+2, \cdots, \ell).$$

The decision of whether the sequence $X_{t_1}, X_{t_2}, \ldots, X_{t_\ell}$ form a Homogenous Compound Poisson Process against the alternative that a change point exists at an unknown interval j,can be expressed in the following hypothesis test.

$$H_0 : (\mu_i, \sigma_i^2, \lambda_i) = (\mu, \sigma^2, \lambda) \text{ for } i = 1, 2, \cdots, \ell) \tag{5.1}$$

versus the alternative that,

$$H_1 : (\mu_i, \sigma_i^2, \lambda_i) = \begin{cases} (\mu_1, \sigma_1^2, \lambda_1) & \text{for } i = 1, 2, \cdots, j-1 \\ (\delta, \sigma^2, \lambda) & \text{for } i = j \\ (\mu_2, \sigma_2^2, \lambda_2) & \text{for } i = j+1, j+2, \cdots, \ell \end{cases} \tag{5.2}$$

While a likelihood ratio and information approach were proposed for the case of known variance, due to difficulties in the properties of the likelihood function, in the case of variance change, a Bayesian approach is proposed for this model. In Jie Chen, et al. (2010)[30], a Bayesian approach was shown to be effective in a similar

case where the variance of the normal variables was assumed to remain constant.

A Bayesian Approach

Bayesian methods for change point analysis have been widely and effectively used for many different problems. Some examples are Sen and Srivastava (1975) [170] and Erdman and Emerson (2008) [45] who worked with changes in the mean of normal models. Ferreira (1975) [46] and Charlton and Troskie (1999) [19] who worked with Regression models. While Ray and Tsay (2002) [163], and Reboul and Benjelloun (2005) [164] proposed Bayesian models for time series data.

For the change point model defined by equation (5.2) the following change point model is proposed. Do to the inclusion of the gamma function in calculating the posterior distribution of the change location, the change point will be confined to $2 \leq j \leq \ell - 1$. The prior distribution of the change location is assumed to be,

$$\pi_0(j) = \begin{cases} \frac{1}{\ell - 2} & \text{if } 2 \leq j \leq \ell - 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

The prior of the variances, $\sigma_1^2, \sigma_2^2, \sigma^2$, are assumed to be,

$$\pi_0(\sigma_1^2, \sigma_2^2, \sigma^2 | j) \propto \frac{1}{\sigma_1^2 \sigma_2^2 \sigma^2} \tag{5.4}$$

and the prior for the means of the normal distributions, μ_1, μ_2, δ, are assumed to be,

$$\pi_0(\mu_1, \mu_2, \delta | \sigma_1^2, \sigma_2^2, \sigma^2, j) \propto \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}\mu_1^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\delta^2} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}\mu_2^2}. \tag{5.5}$$

The following theorem states the main result.

THEOREM 5.1. For the change-point problem specified by Equation (5.2), un-

der the normality assumption of the random sample x_i, the Poisson assumption of N_{t_i}, and the non-informative priors (5.3), (5.4), and (5.5), the posterior probability distribution function of the change point location j is given by,

$$\pi_1^*(j) = \frac{\pi_1'(j)}{\sum_{i=2}^{\ell-1}\pi_1'(i)} \text{ for } j = 2, 3, \ldots, \ell-1, \tag{5.6}$$

where

$$\begin{aligned}\pi_1'(j) &= \frac{a^{\frac{-j+1}{2}} b^{\frac{j-\ell}{2}} c^{-\frac{1}{2}} \Gamma\left(\frac{j-1}{2}\right)\Gamma\left(\frac{\ell-j}{2}\right)}{\sqrt{1+\sum_{i=1}^{j-1} m_i}\sqrt{1+m_j}\sqrt{1+\sum_{i=j+1}^{\ell} m_i}} \\ &\quad \times \left(\frac{\sum_{i=1}^{j-1} m_i}{\sum_{i=1}^{j-1} t_i}\right)^{\sum_{i=1}^{j-1} m_i} \left(\frac{m_j}{t_j}\right)^{m_j} \left(\frac{\sum_{i=j+1}^{\ell} m_i}{\sum_{i=j+1}^{\ell} t_i}\right)^{\sum_{i=j+1}^{\ell} m_i}, \end{aligned} \tag{5.7}$$

and the constants a, b, c are calculated as

$$a = \sum_{i=1}^{j-1} \frac{x_{t_i}^2}{m_i} - \frac{\left(\sum_{i=1}^{j-1} m_i\right)^2}{1+\sum_{i=1}^{j-1} m_i} \tag{5.8}$$

$$b = \sum_{i=j+1}^{\ell} \frac{x_{t_i}^2}{m_i} - \frac{\left(\sum_{i=j+1}^{\ell} m_i\right)^2}{1+\sum_{i=j+1}^{\ell} m_i} \tag{5.9}$$

$$c = \frac{x_{t_j}^2}{m_j(1+m_j)}. \tag{5.10}$$

PROOF. Under the normality assumption of the random sample X_i and the Poisson assumption of N_{t_i}, the likelihood function of the sample under the alternative hypothesis (5.2) can be found as

$$\begin{aligned} L(\mu_1, \mu_2, \delta, \sigma_1, \sigma_2, \sigma, j) &= L(\mu_1, \mu_2, \delta, \sigma_1, \sigma_2, \sigma, j, X_{t_i} | N_{t_i}, i = 1, 2, \cdots, \ell) \\ &\quad \cdot P(N_{t_i} = m_i, i = 1, 2, \cdots, \ell) \end{aligned}$$

$$\propto \quad \frac{1}{(\sigma_1^2)^{\frac{j-1}{2}}} \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{j-1}\left(\frac{x_{t_i}-m_i\mu_1}{\sqrt{m_i}}\right)^2\right\} \frac{1}{(\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{x_{t_j}-m_j\delta}{\sqrt{m_j}}\right)^2\right\}$$

$$\cdot \frac{1}{(\sigma_2^2)^{\frac{\ell-j}{2}}} \exp\left\{-\frac{1}{2\sigma_2^2}\sum_{i=j+1}^{\ell}\left(\frac{x_{t_i}-m_i\mu_2}{\sqrt{m_i}}\right)^2\right\} \cdot P(N_{t_i}=m_i, i=1,2,\cdots,\ell). \quad (5.11)$$

The joint posterior distribution of the parameters $\mu_1, \mu_2, \delta, \sigma_1, \sigma_2, \sigma$ and j can be expressed as proportional to,

$$\pi_1'(\mu_1,\mu_2,\delta,\sigma_1,\sigma_2,\sigma,j) \propto \quad L(\mu_1,\mu_2,\delta,\sigma_1,\sigma_2,\sigma,j,X_{t_i}|N_{t_i}, i=1,2,\cdots,\ell)$$

$$\cdot P(N_{t_i}=m_i, i=1,2,\cdots,\ell)\cdot\pi_0(\mu_1,\mu_2,\delta|\sigma_1^2,\sigma_2^2,\sigma^2,j)$$

$$\cdot\pi_0(\sigma_1^2,\sigma_2^2,\sigma^2|j)\cdot\pi_0(j) \doteq \pi_1(\mu_1,\mu_2,\delta,\sigma_1,\sigma_2,\sigma). \quad (5.12)$$

Where $\pi_0(\cdot)$ are the priors for the parameters as defined above. Now, from priors (5.3)-(5.5) and (5.12), the posterior distribution of j is then proportional to,

$$\pi_1(j) \propto \int_0^\infty\int_0^\infty\int_0^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty \pi_1(\mu_1,\mu_2,\delta,\sigma_1,\sigma_2,\sigma)d\mu_1 d\mu_2 d\delta d\sigma_1^2 d\sigma_2^2 d\sigma^2. \quad (5.13)$$

Now, as the function $\pi_1(\mu_1,\mu_2,\delta,\sigma_1,\sigma_2,\sigma)$ is separable in the integration variable, the integration is given in parts below, beginning with the integration with respect to μ_1. From Equation (5.11),

$$\int_{-\infty}^\infty \pi_1(j)d\mu_1 \propto \int_{-\infty}^\infty \frac{1}{(\sigma_1^2)^{\frac{j-1}{2}}} e^{\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{j-1}\left(\frac{x_{t_i}-m_i\mu_1}{\sqrt{m_i}}\right)^2\right\}} \frac{1}{\sqrt{2\pi}\sigma_1} e^{\left\{-\frac{1}{2\sigma_1^2}\mu_1^2\right\}} d\mu_1. \quad (5.14)$$

Now, let A equal the exponent in Equation (5.14) that is,

$$A \quad = \quad -\frac{1}{2\sigma_1^2}\left\{\sum_{i=1}^{j-1}\left(\frac{x_{t_i}-m_i\mu_1}{\sqrt{m_i}}\right)^2 + \mu_1^2\right\}$$

$$= -\frac{1}{2\sigma_1^2}\left\{\sum_{i=1}^{j-1}\frac{x_{t_i}^2}{m_i} - \frac{\left(\sum_{i=1}^{j-1} x_{t_i}\right)^2}{1+\sum_{i=1}^{j-1} m_i}\right\}$$

$$- \frac{1+\sum_{i=1}^{j-1} m_i}{2\sigma_1^2}\left\{\mu_1 - \frac{\sum_{i=1}^{j-1} x_{t_i}}{1+\sum_{i=1}^{j-1} m_i}\right\}^2. \quad (5.15)$$

By substituting a into the equation as defined by Equation (5.8), Equation (5.15) becomes,

$$A = -\frac{a}{2\sigma_1^2} - \frac{1+\sum_{i=1}^{j-1} m_i}{2\sigma_1^2}\left\{\mu_1 - \frac{\sum_{i=1}^{j-1} x_{t_i}}{1+\sum_{i=1}^{j-1} m_i}\right\}^2. \quad (5.16)$$

Now, by substituting A back into Equation (5.14) and multiplying in the right constant,

$$\int_{-\infty}^{\infty}\pi_1(j)d\mu_1 \propto \left(\frac{1}{\sigma_1^2}\right)^{\frac{j-1}{2}} e^{-\frac{a}{2\sigma_1^2}}\frac{1}{\sqrt{1+\sum_{i=1}^{j-1} m_i}}$$

$$\cdot\int_{-\infty}^{\infty}\frac{\sqrt{1+\sum_{i=1}^{j-1} m_i}}{\sqrt{2\pi}\sigma_1}\cdot e^{-\frac{1+\sum_{i=1}^{j-1} m_i}{2\sigma_1^2}\left\{\mu_1-\frac{\sum_{i=1}^{j-1} x_{t_i}}{1+\sum_{i=1}^{j-1} m_i}\right\}^2} d\mu_1$$

$$= \left(\frac{1}{\sigma_1^2}\right)^{\frac{j-1}{2}} e^{-\frac{a}{2\sigma_1^2}}\frac{1}{\sqrt{1+\sum_{i=1}^{j-1} m_i}}. \quad (5.17)$$

Note: The integra is the integral of a normal variable over its entire domain and thus equals one.

Now, for the integration with respect to μ_2, from Equation (5.11),

$$\int_{-\infty}^{\infty}\pi_1(j)d\mu_2 \propto \int_{-\infty}^{\infty}\frac{1}{(\sigma_1^2)^{\frac{\ell-j}{2}}}e^{\left\{-\frac{1}{2\sigma_2^2}\sum_{i=j+1}^{\ell}\left(\frac{x_{t_i}-m_i\mu_2}{\sqrt{m_i}}\right)^2\right\}}\frac{1}{\sqrt{2\pi}\sigma_2}e^{\left\{-\frac{1}{2\sigma_2^2}\mu_2^2\right\}}d\mu_2. \quad (5.18)$$

Similar to the process for μ_1, let B equal the exponent in Equation (5.18) and it becomes,

$$B = -\frac{1}{2\sigma_2^2}\left\{\sum_{i=j+1}^{\ell}\frac{x_{t_i}^2}{m_i} - \frac{\left(\sum_{i=j+1}^{\ell} x_{t_i}\right)^2}{1+\sum_{i=j+1}^{\ell} m_i}\right\}$$
$$- \frac{1+\sum_{i=j+1}^{\ell} m_i}{2\sigma_1^2}\left\{\mu_2 - \frac{\sum_{i=j+1}^{\ell} x_{t_i}}{1+\sum_{i=j+1}^{\ell} m_i}\right\}^2. \quad (5.19)$$

.

By substituting b into the above equation as defined by Equation (5.9), Equation (5.19) becomes,

$$B = -\frac{b}{2\sigma_2^2} - \frac{1+\sum_{i=j+1}^{\ell} m_i}{2\sigma_2^2}\left\{\mu_2 - \frac{\sum_{i=j+1}^{\ell} x_{t_i}}{1+\sum_{i=j+1}^{\ell} m_i}\right\}^2. \quad (5.20)$$

By substituting B back into Equation (5.18) and multiplying in the right constant,

$$\int_{-\infty}^{\infty}\pi_1(j)d\mu_2 \propto \left(\frac{1}{\sigma_2^2}\right)^{\frac{\ell-j}{2}} e^{-\frac{b}{2\sigma_2^2}}\frac{1}{\sqrt{1+\sum_{i=j+1}^{\ell} m_i}}\cdot\int_{-\infty}^{\infty}\frac{\sqrt{1+\sum_{i=j+1}^{\ell} m_i}}{\sqrt{2\pi}\sigma_2}$$
$$\cdot e^{-\frac{1+\sum_{i=j+1}^{\ell} m_i}{2\sigma_2^2}\left\{\mu_2 - \frac{\sum_{i=j+1}^{\ell} x_{t_i}}{1+\sum_{i=j+1}^{\ell} m_i}\right\}^2} d\mu_2$$
$$= \left(\frac{1}{\sigma_2^2}\right)^{\frac{j-1}{2}} e^{-\frac{b}{2\sigma_2^2}}\frac{1}{\sqrt{1+\sum_{i=j+1}^{\ell} m_i}}. \quad (5.21)$$

Step 3, the integration with respect to δ, from Equation (5.11),

$$\int_{-\infty}^{\infty}\pi_1(j)d\delta \propto \int_{-\infty}^{\infty}\frac{1}{(\sigma^2)^{\frac{1}{2}}}\cdot\exp\left\{-\frac{1}{2\sigma^2}\frac{\left(x_{t_j}-m_j\delta\right)^2}{m_j}\right\}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{\left\{-\frac{1}{2\sigma^2}\delta^2\right\}}d\delta. \quad (5.22)$$

Letting C be the exponential portion of the integrand and,

$$C = -\frac{1}{2\sigma^2}\frac{\left(x_{t_j}-m_j\delta\right)^2}{m_j} - \frac{1}{2\sigma^2}\delta^2$$

$$= \quad -\frac{1}{2\sigma^2}\frac{x_{t_j}^2}{m_j(1+m_j)} - \frac{1+m_j}{2\sigma^2}\left[\delta - \frac{x_{t_j}}{1+m_j}\right]^2. \tag{5.23}$$

Now, substitute c from Equation (5.10) and C becomes,

$$C = -\frac{c}{2\sigma^2} - \frac{1+m_j}{2\sigma^2}\left[\delta - \frac{x_{t_j}}{1+m_j}\right]^2. \tag{5.24}$$

By substituting C back into Equation (5.22) and multiplying in the right constant,

$$\begin{aligned}\int_{-\infty}^{\infty}\pi_1(j)d\delta \quad &\propto \quad \frac{1}{(\sigma^2)^{\frac{1}{2}}}\cdot e^{-\frac{c}{2\sigma^2}}\frac{1}{\sqrt{1+m_j}} \\ &\qquad \cdot\int_{-\infty}^{\infty}\frac{\sqrt{1+m_j}}{\sqrt{2\pi}\sigma}e^{-\frac{1+m_j}{2\sigma^2}\left[\delta-\frac{x_{t_j}}{1+m_j}\right]^2}d\delta \\ &= \quad \frac{1}{(\sigma^2)^{\frac{1}{2}}}\cdot e^{-\frac{c}{2\sigma^2}}\frac{1}{\sqrt{1+m_j}}.\end{aligned} \tag{5.25}$$

Now of the three previous results, only the result from integrating μ_1 contained σ_1^2, so beginning with Equation (5.17),

$$\begin{aligned}\int_0^{\infty}\pi_1(j)d\sigma_1^2 \quad &\propto \quad \int_0^{\infty}\left(\frac{1}{\sigma_1^2}\right)^{\frac{j-1}{2}}\frac{1}{\sqrt{1+\sum_{i=1}^{j-1}m_i}}e^{-\frac{a}{2\sigma_1^2}}\frac{1}{\sigma_1^2}d\sigma_1^2 \\ &= \quad \int_0^{\infty}\frac{1}{\sqrt{1+\sum_{i=1}^{j-1}m_i}}\left(\sigma_1^2\right)^{\frac{-j-1}{2}}e^{-\frac{a}{2\sigma_1^2}}d\sigma_1^2 \\ &= \quad \frac{1}{\sqrt{1+\sum_{i=1}^{j-1}m_i}}\int_0^{\infty}\left(\sigma_1^2\right)^{\frac{-j-1}{2}}e^{-\frac{a}{2\sigma_1^2}}d\sigma_1^2.\end{aligned} \tag{5.26}$$

Let $\frac{a}{2\sigma_1^2} = u$, then $d\sigma_1^2 = -\frac{a}{2u^2}du$, substituting into the equation above,

$$= \quad \frac{1}{\sqrt{1+\sum_{i=1}^{j-1}m_i}}\left[-\int_{\infty}^{0}\left(\frac{a}{2u}\right)^{\frac{-j-1}{2}}e^{-u}\frac{a}{2u^2}du\right]$$

$$= \frac{\left(\frac{a}{2}\right)^{\frac{-j+1}{2}}}{\sqrt{1+\sum_{i=1}^{j-1} m_i}} \int_0^\infty u^{\left(\frac{j-1}{2}\right)-1} e^{-u} du$$

$$\propto \frac{(a)^{\frac{1-j}{2}}}{\sqrt{1+\sum_{i=1}^{j-1} m_i}} \cdot \Gamma\left(\frac{j-1}{2}\right), \tag{5.27}$$

provided $\frac{j-1}{2} > 0$, or $j > 1$.

Similar to the integration for σ_1^2, σ_2^2 was confined to Equation (5.21) so,

$$\int_0^\infty \pi_1(j) d\sigma_2^2 \propto \int_0^\infty \left(\frac{1}{\sigma_2^2}\right)^{\frac{\ell-j}{2}} \frac{1}{\sqrt{1+\sum_{i=j+1}^{\ell} m_i}} e^{-\frac{b}{2\sigma_2^2}} \frac{1}{\sigma_2^2} d\sigma_2^2$$

$$= \frac{1}{\sqrt{1+\sum_{i=j+1}^{\ell} m_i}} \int_0^\infty \left(\sigma_2^2\right)^{\frac{-\ell+j-2}{2}} e^{-\frac{b}{2\sigma_2^2}} d\sigma_2^2$$

$$\propto \frac{(b)^{\frac{j-\ell}{2}}}{\sqrt{1+\sum_{i=j+1}^{\ell} m_i}} \cdot \Gamma\left(\frac{\ell-j}{2}\right), \tag{5.28}$$

provided $\frac{\ell-j}{2} > 0$, or $j < \ell$.

Finally, σ^2 was contained in Equation (5.25) and,

$$\int_0^\infty \pi_1(j) d\sigma^2 \propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} \frac{1}{\sqrt{1+m_j}} e^{-\frac{c}{2\sigma^2}} \frac{1}{\sigma^2} d\sigma^2$$

$$= \frac{1}{\sqrt{1+m_j}} \int_0^\infty \left(\sigma^2\right)^{-\frac{3}{2}} e^{-\frac{c}{2\sigma^2}} d\sigma^2.$$

$$= \frac{\left(\frac{c}{2}\right)^{-\frac{1}{2}}}{\sqrt{1+m_j}} \cdot \Gamma\left(\frac{1}{2}\right)$$

$$\propto \frac{c^{-\frac{1}{2}}}{\sqrt{1+m_j}}. \tag{5.29}$$

Combining Equations (5.27),(5.28), and (5.29),

$$\pi_1(j) \propto \frac{(a)^{\frac{1-j}{2}}}{\sqrt{1+\sum_{i=1}^{j-1} m_i}} \cdot \frac{c^{-\frac{1}{2}}}{\sqrt{1+m_j}} \cdot \frac{(b)^{\frac{j-\ell}{2}}}{\sqrt{1+\sum_{i=j+1}^{\ell} m_i}} \cdot \Gamma\left(\frac{j-1}{2}\right)$$

$$\cdot\Gamma\left(\frac{\ell-j}{2}\right) \cdot P\left(N_{t_i} = m_i, i = 1, 2, \cdots, \ell\right). \tag{5.30}$$

Further, as

$N_{t_i} \sim Poisson(\lambda_1 t_i)$ for $(i = 1, 2, \cdots, j-1)$

$N_{t_i} \sim Poisson(\lambda t_i)$ for $(i = j)$

$N_{t_i} \sim Poisson(\lambda_2 t_i)$ for $(i = j+1, j+2, \cdots, \ell)$,

$$P\left(N_{t_i} = m_i, i = 1, 2, \cdots, \ell\right) = \prod_{i=1}^{\ell} P\left(N_{t_i} = m_i\right) \tag{5.31}$$

$$= \left(\prod_{i=1}^{j-1} \frac{(\lambda_1 t_i)^{m_i} e^{-\lambda_1 t_i}}{m_i!}\right) \left(\frac{(\lambda t_j)^{m_j} e^{-\lambda t_j}}{m_j!}\right) \left(\prod_{i=j+1}^{\ell} \frac{(\lambda_2 t_i)^{m_i} e^{-\lambda_2 t_i}}{m_i!}\right)$$

$$= \frac{\lambda_1^{\sum_{i=1}^{j-1} m_i} \prod_{i=j+1}^{\ell} t_i^{m_i} e^{-\lambda_1 \sum_{i=1}^{j-1} t_i}}{\prod_{i=1}^{j-1} m_i!} \left(\frac{(\lambda t_j)^{m_j} e^{-\lambda t_j}}{m_j!}\right) \frac{\lambda_2^{\sum_{i=j+1}^{\ell} m_i} \prod_{i=j+1}^{\ell} t_i^{m_i} e^{-\lambda_2 \sum_{i=j+1}^{\ell} t_i}}{\prod_{i=j+1}^{\ell} m_i!}$$

$$= \frac{\lambda_1^{\sum_{i=1}^{j-1} m_i} \prod_{i=j+1}^{\ell} t_i^{m_i}}{\prod_{i=1}^{j-1} m_i!} \left(\frac{(\lambda t_j)^{m_j}}{m_j!}\right) \frac{\lambda_2^{\sum_{i=j+1}^{\ell} m_i} \prod_{i=j+1}^{\ell} t_i^{m_i}}{\prod_{i=j+1}^{\ell} m_i!}$$

$$\cdot e^{-\lambda_1 \sum_{i=1}^{j-1} t_i - \lambda t_j - \lambda_2 \sum_{i=j+1}^{\ell} t_i}. \tag{5.32}$$

In order to calculate the probability we need to find the MLE's of the lambda's this can easily be done an shall not be reproduced as the details are similar to those in Chapter 4. The estimators are,

$$\hat{\lambda_1} = \frac{\sum_{i=1}^{j-1} m_i}{\sum_{i=1}^{j-1} t_i}, \quad \hat{\lambda_1} = \frac{m_j}{t_j}, \quad \hat{\lambda_2} = \frac{\sum_{i=j+1}^{\ell} m_i}{\sum_{i=j+1}^{\ell} t_i}. \tag{5.33}$$

Substituting them into Equation (5.32), the estimated $P\left(N_{t_i} = m_i, i = 1, 2, \cdots, \ell\right)$

is,

$$= \frac{\prod_{i=1}^{\ell} t_i^{m_i}}{\prod_{i=1}^{\ell} m_i!} \cdot e^{-\sum_{i=1}^{\ell} m_i} \left(\frac{\sum_{i=1}^{j-1} m_i}{\sum_{i=1}^{j-1} t_i} \right)^{\sum_{i=1}^{j-1} m_i} \left(\frac{m_j}{t_j} \right)^{m_j} \left(\frac{\sum_{i=j+1}^{\ell} m_i}{\sum_{i=j+1}^{\ell} t_i} \right)^{\sum_{i=j+1}^{\ell} m_i} . \quad (5.34)$$

As the first two factors of (5.34) are constant for all j, we only include the last three factors in $\pi_1(j)$. Substitution the last three factors of (5.34) into (5.30) and the proof of Theorem 5.1 is complete. □

A Bayesian approach to locating a single change point, j. is given by $\hat{j}$ such that,

$$\pi_1^* \hat{j} = \max_{2 \le j \le \ell - 1} \pi_1(j). \quad (5.35)$$

Simulation Study

As with the methods in Chapter 4, once the Bayesian method was developed, an extensive simulation study was performed using R. The simulation was run on data with change in both mean, μ, and variance, σ^2 of the normal variable along with a change in the mean parameter of the Poisson λ. For purposes of simulation, the number of observations in each interval, m_i, was set equal to one, and the distances, t_i's were generated using the exponential distribution.

In the study, three sets of simulations were run for different levels of change. In each of the three sets, $\mu_1 = \delta = 0$, $\sigma_1^2 = \sigma^2 = 1$, and $\lambda_1 = \lambda = 1$. The low level of change was $\mu_2 = 3$, $\sigma_2^2 = 1.5$, and $\lambda_2 = 2$, while the medium level of change was $\mu_2 = 5$, $\sigma_2^2 = 2$, and $\lambda_2 = 3$, and the large level of change was $\mu_2 = 7$, $\sigma_2^2 = 2.5$, and $\lambda_2 = 5$. For each level of change, the sample sizes were set as $n = 40, 75, 100, 125,$ and 200, and the location of the change was placed at the $\frac{n}{4}th$, $\frac{n}{2}$th, and $\frac{3n}{4}$th

Table 8. Bayesian Simulation with n=40

$(\mu_2, \sigma_2, \lambda_2)$	Position	f	$\bar{j^*}$	$MSE(j^*)$
	10	0.655	9.918	2.022
$(3, 1.5, 2)$	20	0.660	19.892	1.098
	30	0.595	29.635	2.439
	10	0.842	10.020	0.196
$(5, 1, 3)$	20	0.859	20.004	0.216
	30	0.791	29.910	0.480
	10	0.927	10.049	0.073
$(7, 2.5, 5)$	20	0.926	20.046	0.074
	30	0.901	30.011	0.139

positions in the sequence. For each combination of level of change, change location, and sample size, 1000 repetitions were performed and the frequency of selecting the correct location j was recorded along with the estimated mean square error, denoted $\text{MSE}(j^*)$, of the position of change and the average of the estimated change location, denoted $\bar{j^*}$. The results are given in Table 8 through Table 12.

Table 9. Bayesian Simulation with n=75

$(\mu_2, \sigma_2, \lambda_2)$	Position	f	$\bar{j^*}$	$MSE(j^*)$
	19	0.686	18.973	0.875
$(3, 1.5, 2)$	38	0.666	37.998	0.802
	56	0.595	55.903	0.883
	19	0.864	19.067	0.157
$(5, 1, 3)$	38	0.878	38.018	0.140
	56	0.861	56.018	0.186
	19	0.935	19.057	0.065
$(7, 2.5, 5)$	38	0.923	38.063	0.077
	56	0.922	56.049	0.087

The simulations show that the method is effective in locating the position of the change even at the low level of change. As would be expected, at the low and middle levels of change, method improved in finding the location of the change as the sample sized increases.This improvement is not seen in the case of a large level of change. At the lower level of change, the method successfully located the change 65% to 70%, while finding 84% to 88% of the time for the medium level of change, and 91.5% to 93.6% for the large level of change. The MSE(j^*)'s for all cases would indicate that the method is proficient at locating the position of change with the values being less than one, except for the case of low level of change with sample size 40, in which case

Table 10. Bayesian Simulation with n=100

$(\mu_2, \sigma_2, \lambda_2)$	Position	f	$\bar{j^*}$	$MSE(j^*)$
	25	0.704	24.978	0.838
$(3, 1.5, 2)$	50	0.696	49.970	0.832
	75	0.676	74.932	0.960
	25	0.874	25.073	0.155
$(5, 1, 3)$	50	0.863	50.063	0.137
	75	0.866	75.034	0.184
	25	0.914	25.083	0.089
$(7, 2.5, 5)$	50	0.920	50.070	0.086
	75	0.922	75.049	0.081

the MSE(j^*)'s ranged from 2.5 and 1.1. The Average estimated location also suggest a good estimation as they are all within two tenths of a unit of the actual position. It should be noted that the level of change in the variance is smaller than in the other variables. While the model is supposed to identify changes in variance, large changes in variance on the same magnitude as the changes in the other variables, seems to confound the process and the percentage of location of change drops significantly.

In addition to the data sets described above, null sets, with no change were simulated for $n = 40$ and 200. The value of the parameters were based on the values found in aCGH data from the next section. One question that must be addressed in

Table 11. Bayesian Simulation with n=125

$(\mu_2, \sigma_2, \lambda_2)$	Position	f	$\bar{j^*}$	$MSE(j^*)$
	31	0.709	31.000	0.624
$(3, 1.5, 2)$	63	0.674	62.937	0.733
	94	0.661	93.936	0.864
	31	0.878	31.084	0.140
$(5, 1, 3)$	63	0.880	63.062	0.132
	94	0.850	94.060	0.164
	31	0.916	31.078	0.084
$(7, 2.5, 5)$	63	0.936	63.052	0.064
	94	0.922	94.043	0.081

applying this Bayesian method is what threshold should be used to determine that a change has taken place. Simulations indicate that a high threshold of, 0.8 or 0.9, should be set as indicated by the results in Table 13.

Table 12. Bayesian Simulation with n=200

$(\mu_2, \sigma_2, \lambda_2)$	Position	f	$\bar{j^*}$	$MSE(j^*)$
	50	0.704	50.012	0.606
$(3, 1.5, 2)$	100	0.709	100.025	0.731
	150	0.668	2149.946	0.900
	50	0.873	50.100	0.136
$(5, 1, 3)$	100	0.879	100.0844	0.138
	150	0.874	150.047	0.153
	50	0.923	50.077	0.077
$(7, 2.5, 5)$	100	0.931	100.059	0.069
	150	0.915	150.071	0.091

Table 13. False Positive Rates

	Sample Size	
Threshold	40	200
0.5	0.335	0.264
0.6	0.229	0.176
0.7	0.151	0.118
0.8	0.084	0.074
0.9	0.042	0.046

Application to aCGH Data

The Bayesian method developed was applied to the same nine cell lines described in the previous chapter. One thing that was not noted in the description of the data preparation in the previous chapter was the combining of intervals with a $Log_2(G_i/T_i) = 0$ with the previous interval. This is necessary because the power of c is negative, which causes a division by zero. Another technical issue with this method is the number of intervals to be tested. Due to the presence of the gamma function in the calculation of the posterior distribution the sample cannot exceed 300 intervals. Therefore, for the single change point model to be applied to an entire cell line data points must be combined.

While the Bayesian method developed is for a single change point, unlike the application of the two previous methods, a sliding window approach for identifying multiple changes in location will be presented, and applied to five chromosomes. This is an important feature as one can see from the earlier applications that the occurrence of multiple change points on a chromosome can be common. There is also another application for the multiple change point models, which is the case where the copy number variation occurs across an entire chromosome. This later case will be demonstrated with cell line GM00143, in which the multiple change procedure is used.

With a sliding window there are a few technical issues that must be worked out. The first is the problem with edge effect, where the posterior distribution can not be calculated for the first and last positions. The need to overlap the windows by a point or two is well documented in, Sun et al. (2006) [186], and Chen et al.(2010)[30].

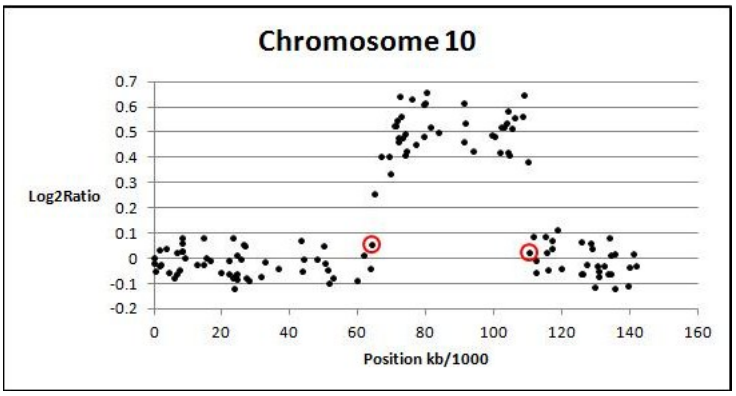

Figure 5. GM05296 Chromosome 10

In this case it requires two points of overlap to insure that posterior probability is calculated for each interior point. Another issue is what window size to use. In Sun et al. (2006)[186] a range of 3 to 10 consecutive markers was recommended, while in Chen et al.(2010)[30] used a range of window sizes from 12 to 35. In this application a range of window sized from 10 to 25 are used for a chromosome. As indicated in the simulation study, the threshold for the test needs to be high, in the range of 0.8 to 0.9. The method proposed is modeled after that in Chen et al. (2010)[30]. Another issue is what to do with the end of the data set. For the last window, the size may be larger or smaller than the set size. If the last window size would be less than .3 of the window size, the points will be combined with, what would have been the next to last window, which will be slightly larger than the set size, and if larger than .3 times the window size the points will be evaluated in their own window. This is numerically necessary as the minimum number of points must be greater than three. These extremely small data sets may also artificially create a point with a high posterior probability. In this application, a heuristic threshold of 0.5 for the

maximum posterior probabilities was used.

The steps to the algorithm are as follows.

1. If a chromosome has only one suspected change location, then apply the single change point model proposed, calculate the posterior probabilities for each location by Equation 5.7 and applying Theorem 5.1, then use Equation 5.35 to find the estimated location of change.
2. If multiple changes in a chromosome are suspected, break the chromosome into windows starting with size 10, with two points overlapping each window.
3. For each window, calculate the posterior probability using Equation 5.7 and applying Theorem 5.1, then use Equation 5.35 to find the estimated location of change.
4. Record the location of change in each window along with its posterior probability.
5. Repeat steps 2 - 4 for window sizes 11, 12, $\cdots$, J, where J is the window size in which the posterior probabilities and positions seem to stabilize.
6. Determine from the last results which of the maximum posterior probabilities exceed the threshold and are deemed actual change points.

The results of the application indicate that the maximum posterior does a good job at identifying the location of change in an interval where one exists. However, since some of the intervals have extremely small or large lengths, in intervals with no change in the copy number the clustering of points are enough to sway the posterior probability. In other words, the method does what it is supposed to by being sensitive to extreme changes in λ as well as μ. This problem seemed to diminish with the

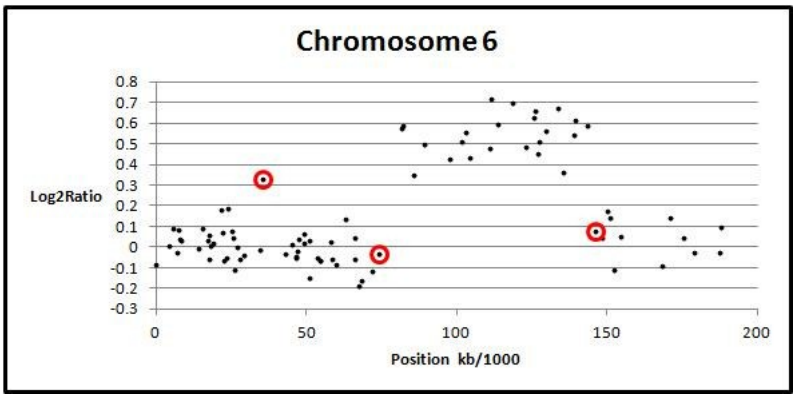

Figure 6. GM01524 Chromosome 6

implementation of the sliding window. However, due to the fact that many of the first genes at the beginning of the chromosome were in the zero position and given a length of one, this leads to the posterior probability of position two on the chromosome to be extremely high, in the first window. The results presented here have dismissed these false positive as an error in setting up the data. Figure 5, shows Chromosome 10 of cell line GM05296, the red rings indicate the identified change points throughout the graphs in this section.

In Table 14, the results of the sliding window method are given. As a comparison, Table 15 shows the results of some of the chromosomes with one change point. As can be seen, the maximum posterior probabilities are what we would expect using the sliding window, and lower than we would expect for the single change model. Also, the false positive rate is greatly improved using the sliding window algorithm.

The graphs of the data for the rest of the two change cases are listed throughout this section for review. It is worth noting that one of the false positives from

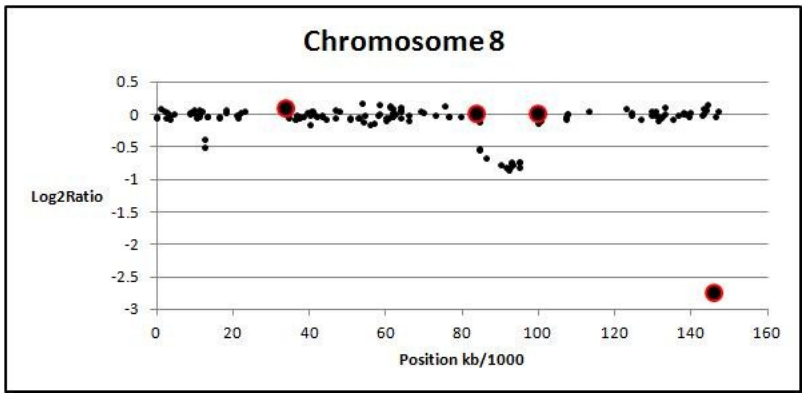

Figure 7. GM03134 Chromosome 8

Chromosome 8 of cell line GM03134 as well as the one from Chromosome 6 of cell line GM01524 are outliers, see Figures 6 and 7. Chromosome 11 of cell line GM05296 was the chromosome the method performed the worst on. Note in In Figure 9, the red square indicates a known change point which was missed as well as three false positives. While the missed location was chosen for smaller window sizes, the posterior probability for each actual change location was low, see Table 14 Gene Marker CTD-2208j5. To try and improve the results for this chromosome, a second analysis was completed. This time, adjacent data points were combined to achieve a minimum interval length of 100 kb; however, the results were not significantly improved. In the second run, the posterior probability increased for the missed change points; however it still failed to exceed the threshold of 0.5, while the false positives increased to 7. Due to time restrictions, no further analysis of this chromosome was performed.

While the single change point model seems inconsistent in its performance with aCGH data, the sliding window method seems to be a viable method for finding the location of copy number variations in aCGH data.

Table 14. Results for Chromosomes with Two Copy Number Changes

Cell Line	Chromosome	Gene Marker	$\pi_1(\hat{j})$	Window size	FP
GM01524	6	CTD-2009c06	0.82283	21	1
	6	RP11-139o22	0.72919	21	
GM03134	8	RP11-107F03	0.81935	25	2
	8	RP11-102K07	0.67581	25	
GM05296	10	RP11-14i14	0.78348	25	0
	10	RP11-46g13	0.68266	25	
GM05296	11	CTD-2208j5	0.27164	13	3
	11	RP11-18B09	0.70956	13	
GM13031	17	RP5-1071i14	0.7161	21	2
	17	RP11-670E13	0.91776	21	

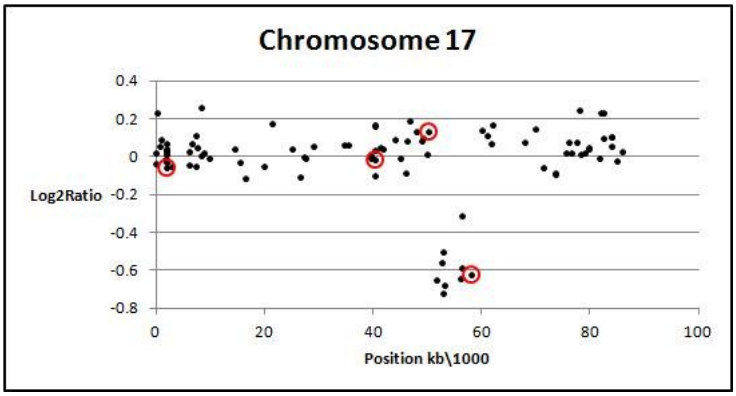

Figure 8. GM13031 Chromosome 17

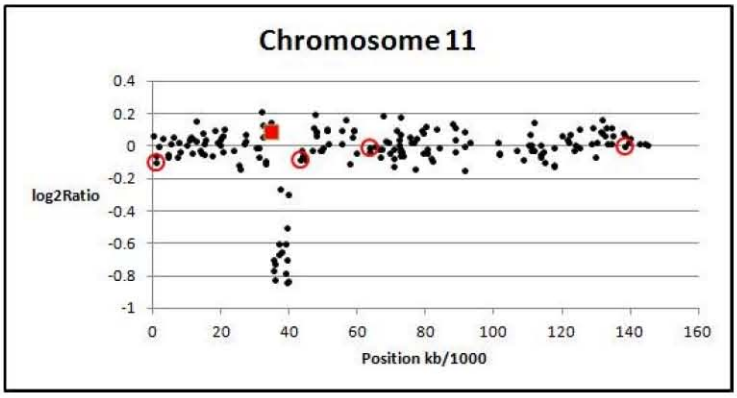

Figure 9. GM05296 Chromosome 11

Table 15. Results for Chromosomes with One Copy Number Changes

Cell Line	Chromosome	Gene Marker	$\pi_1(\hat{j})$	FP
GM01535	12	RP11-81g12	0.711908	8
GM01750	9	RP11-85J05	0.82191	9
	14	RP11-48l01	0.59438	
GM03563	3	CTD-2014B13	0.907387	5
	9	RP11-28n06	0.994078	

CHAPTER 6

FUTURE WORK

While the Bayesian model presented in Chapter 5, meets the goal that was set in chapter 3, to identify multiples changes in all three model parameters, there is clearly much more that can be done. It has become evident that the more a problem is studied and the more one learns about a topic, the more they understand how little they truly know. Each of the three methods presented can use refinement in one aspect or another.

1. For the Likelihood method to be useful, the null distribution must be derived. This is clearly the most challenging and likely to be a long-term goal, as the normal and Poisson variables cannot be separated.
2. The likelihood method could be extended to include change in variance.
3. The completion of one will also provide the possibility for deriving a test statistic for the information criterion, which could lead to a p=value for the test.
4. Until 3 can be done, the method could be extended to test for multiple changes.
5. More extensive work can be done to refine the slide window algorithm and improve its performance. This would include further analysis of Chromosome 11 of GM05296.
6. The Bayesian method can be refined to reduce the false positive rate.

Of these the last is the goal for the near future, while the first could take a lifetime of work, as the exact null distribution for a change point problem under any distribution is still unknown in the literature. Finally, as a future work, I would like to apply this or a similar method in a study of the stock market to determine if there have been structural changes in the growth of the market due to changing the nature of investing.

REFERENCES

1. Hirotogu Akaike, *Information theory and an extention of the maximum likelihood principle*, 2nd International Symposium of Information Theory,B.N. Petrov and E. Csaki, eds., Akademiai Kiado, Budapest, 267–281.
2. Adel Alaeddini, Mehdi Ghazanfari, and Majid Amin Nayeri, *A hybrid fuzzy-statistical clustering approach for estimating the time of changes in fixed and variable sampling control charts*, Information Sciences **179** (2009), 1769–1784.
3. Paul S. Albert, Sally A. Hunsberger, Nan Hu, and Philip R. Taylor, *Identifying multiple changepoints in heterogeneous binary data with an application to molecular genetics*, Biostatistics **5** (2004), 515–529.
4. Anestis Antoniadis, and Irène Gijbels, *Detecting abrupt changes by wavelet methods*, Nonparametric Statistics **14** (2002), 7–29.
5. , Jaromír Antoch, Marie Hušková, Alicja Janic, and Teresa Ledwina, *Data driven rank test for the change point problem*, Metrika **68** (2008), 1–15.
6. Srikesh G. Arunajadai *A point process driven multiple change point model: a robust resistant approach*, Mathematical Biosciences **220** (2009), 57–71.
7. Alexander Aue, and Lojos Horváth, *Delay time in sequenctial detection of change*, Statistics & Probablility Letters **67** (2004), 221–231.
8. Jushan Bai, and Pierre Perron, *Estimating and testing linear models with multiple structural changes*, Econometrica **66** (1998), 47–78.
9. Daniel Barry, and J.A. Hartigan, *Product partition models for change point prob-*

lems, The Annals of Statistics **20** (1992), 260–279.

10. Erhan Bayraktar, and Semith Sezer, *Online change deteection for a Poisson process with a phase-type change-time prior distribution*, Sequential Analysis **28** (2009), 218–250.
11. G.K. Bhattacharyya, and Richard A. Johnson, *Nonparametric tests for shift at an unknown time point*, The Annals of Mathematical Statistics **39** (1968), 1731–1743.
12. Sven Bilke, Qing-Rong Chen, Craig C. Whiteford, and Javed Khan, *Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays*, Bioinformatics **21** (2005), 1138–1145.
13. Wolfgang Bischoff, Enkelejd Hashorva, Jürg Hüsler, and Frank Miller, *On the power of the Kolmogorov test to detect the trend of a Brownian bridge with applications to a change-point problem in regression models*, Statistics & Probability Letters **66** (2004), 105–115.
14. Phelippe Brot and Sylvia Richardson, *Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model*, Bioinformatics **22** (2006), 911–918.
15. Marlo Brown, and Shelemyahu Zacks, *A note on optimal stopping for possible change in the intensity of an ordinary Poisson process*, Statistics & Probability Letters **76** (2006), 1417–1425.
16. Marlo Brown, *Bayesian detection of changes of a poisson process monitored at discrete time points where the arrival rates are unknown*, Sequential Analysis **27** (2008), 68–77.

17. Edward Bryden, and John B. Carlson, *On disinflation since 1982: an application of change-point tests*, Economic Review - Federal Reserve Bank of Cleveland **30** (1994), 31–42.

18. Bradley P. Carlin, Alan E. Gelfand, and Adrian F.M. Smith, *Hierarchical Bayesian analysis of changepoint problems*, Applied Statistics **41** (1992), 389–405.

19. D.O. Chalton, and C.G. Troskie, *Parameter changes in the multiples regression model with autocorrelated erros: Bayesian analysis*, Communications in Statistics - Theory and Methods **28** (1999), 137–142.

20. Florent Chatelain, Jean-Yves Tourneret, Jordi Inglada, and André Ferrari, *Bivariate gamma distributions for image registration and change detection*, IEEE Transactions on Image Processing **16** (2007), 1796–1806.

21. Florent Chatelain, Jean-Yves Tourneret, and Jordi Inglada, *Change Detection in multisensor SAR images using bivariate gamma distributions*, IEEE Transactions on Image Processing **17** (2008), 249–258.

22. Jiahua Chen, A.K. Gupta, and Jianmin Pan, *Information criterion and change point problem for regular models*, Sankhyā **68** (2006), 252–282.

23. Jie Chen, *Testing for a change point in linear regression models*, Communications in Statistics - Theory and Methods **27** (1998), 2481–2493.

24. Jie Chen, and A. K. Gupta, *Testing and locating variance changepoints with application to stock prices*, Journal of the American Statistical Association **92** (1997), 739–747.

25. Jie Chen, and A. K. Gupta, *Change point analysis of a Gaussian model*, Statistics

Papers **40** (1999), 323–333.

26. Jie Chen, and A. K. Gupta, *Parametric statistical change point analysis*, Birkhauser, Boston, 2000.
27. Jie Chen, and Arjun K. Gupta, *Parametric statistical change point analysis*, 2nd ed., Birkhauser, Boston, 2012.
28. Jie Chen, and Yu-Ping Wang, *Detection of DNA copy number changes using statistical change-point analysis*, IEEE International Workshop on GENSIPS 2006, 11-12.
29. Jie Chen, and Yu-Ping Wang, *A statistical change point model approach for the detection of DNA copy number variations in array CGH data*, IEEEACM Transactions on Computational Biology and Bioinformatics **6** (2009), 529–541.
30. Jie Chen, Ayten Yiğiter, Yu-Ping Wang, and Hong-Wen Deng, *A Bayesiand analysis for identifying DNA copy number variations using a compound Poisson process*, EURASIP Journal of Bioinformatics and Systems Biology **2010** (2010), Article ID 268513, 10 pages, doi:10.1155/2010/268513.
31. Kuo-mei Chen, Arthur Cohen, Harold Sackrowitz, *Consistent multiple testing for change point*, Journal of Multivariate Analysis **102** (2011), 1339–1343.
32. Peng-An Chen, Hsiao-Fei Liu, and Kun-Mao Chao, *CNVDetector: locating copy number variations using array CGH data*, Bioinformatics **24** (2008), 2773–2775.
33. Ming-Yen Cheng, and Marc Raimondo, *Kernel methods for optimal change-points estimation in derivatives*, Journal of Computational and graphical Sttistics **17** (2008), 56–75.
34. Tsung-Lin Cheng, *An efficient algorithm for estimating a change-point*, Statistics

and Probability Letters **79** (2009), 559–565.

35. H. Chernoff, and S. Zacks, *Estimating the current mean of a normal distribution which is subjected to changes in time*, The Annals of Mathematical Statistics **35** (1964), 999–1018.
36. J.H. Chin Choy, and L.D. Broemeling, *Some Bayesian inferences for a changing linear model*, Technometrics **22** (1980), 71–78.
37. Grace Chiu, Richard Lockhart, and Richard Routledge, *Asymptotic theory for bent-cable regression–basic case*, Journal of Statistical Planning and Inference **127** (2005), 143–156.
38. Gabriela Ciuperca, *The M-estimation in a multi-phase random nonlinear model*, Statistics and Problbility Letters **79** (2009), 573–580.
39. Stephen V. Crowder, *A simple method for studying run-length distributions of exponentially weighted moving average charts*, Technometrics **29** (1987), 401–407.
40. Raoul-Sam Daruwala, Archisman Rudra, Harry Ostrer, Robert Lucito, Michael Wigler, and Bud Mishra, *A versatile statistical analysis algorithm to detect genome copy number variation*, Proc. Natl. Acad. Sci. USA **101**, 16292–16297.
41. Richard A. Davis, Thomas C.M. Lee, and Gabriel A. Rodreguez-Yam, *Break detection for a class on nonlinear time series models*, Journal of Time Series Analysis **29** (2008), 834–867.
42. Keyue Ding, *A lower confidence bound for the change point after a sequential cusum test*, Journal of Statistical Planning and Inference **115** (2003), 311–326.
43. Jan G. De Gooijer, *Detecting change-points in multidimensional stochastic processes*, Computational Statistics & Data Analysis **51** (2006), 1892–1903.

44. Jean-Francois Dupuy, *Estimation in a change-point hazard regression model*, Statistics & Probability Letters **76** (2006), 182–190.

45. Chandra Erdman, and John W. Emerson, *A fast bayesian change point analysis for the segmentation of microarray data*, Bioinformatics **24** (2008), 2143–2148.

46. Pedro E. Ferreira, *A Bayesuab analysis of a switching regression model: known number of regimes*, Journal of the American Statistical Association **70** (1975), 370–374.

47. Stergios B. Fotopoulos, *The geometric convergence rate of the classical change-point estimate*, Statistics and Probability Letters **79** (2009), 131–137.

48. Stergios B. Fotopoulos, and Venkata K. Jandhuala, *On Hinkley's estimator: Inference about the change point*, Statistics & Probability Letters **77** (2007), 1449–1458.

49. Stergios B. Fotopoulos, Venkata K. Jandhyala, and Li Tan, *Asymptotic study of the change-point mle in multivariate Gaussian families under contiguous alternatives*, Journal of Statistical Planning and Inference **139** (2009), 1190–1202.

50. Donald G. Freeman, *Change-point analysis of the growth effects of state banking degregulation*, Economic Inquiry **43** (2005), 601–613.

51. Marianne Frisén, *Optimal sequential surveillance for finance, public health, and other areas*, Sequential Analysis **28** (2009), 310–337.

52. Yun-Xin Fu, and N. Curnow, *Maximum likelihood estimation of multiple change points*, Biometrica **77** (1990), 563–573.

53. Xiaoli Gao, and Jian Huang, *A robust penzlized method for the analysis of noisy DNA copy number data*,BMC Genomics **11** (2010), 517–526.

http://www.biomedcentral.com/1471-2164/11/517, doi: 10.1186/1471-2164-11-517.

54. L. A. Gardner, Jr., *On detecting changes in the mean of normal variates*, The Annals of Mathematical Statistics **40** (1969), 116–126.
55. Ryan Gill, and Michael Baron, *Consistent estimation in generalized broken-line regression*, Journal of Statistical Planning and Inference **126** (2004), 441–460.
56. Amrit L. Goel, and S.M. Wu, *Determination of A. R. L. and a contour nomogram for cusum charts to control normal mean*, Technometrics **13** (1971), 221–230.
57. Edit Gombay, *Change detection in autoregressive time series*, Journal of Multivariate Analysis **99** (2008), 451–464.
58. Edit Gombay, and Lojos Horváth, *Rates of convergence for U-statistic processes and their bootstrapped versions*, Journal of Statistical Planning and Inference **102** (2002), 247–272.
59. Gérard Grégoire, and Zouhir Hamrouni, *Two non-parametric tests for change-point problems*, Nonparametric Statistics **14** (2002), 87–112.
60. Gérard Grégoire, and Zouhir Hamrouni, *Change point estimation by local linear smoothing*, Journal of Multivariate Analysis **83** (2002), 56–83.
61. A. K. Gupta, and Jie Chen, *Detection changes of mean in multidimentional normal sequences with applications to literature and geology*, Computational Statistics **11** (1996), 211–221.
62. A. K. Gupta, and Asoka Ramanayake, *Change points with linear trend for the exponential distribution*, Journal of Statistical Planning and Inference **93** (2001), 181–195.

63. Gregory Gurevich, and Albert Vexler, *Chagne point problems in the model of logistic regression*, Journal of Statistical Planning and Inference **131** (2005), 313–331.

64. Gregory Gurevich, and Albert Vexler, *Guaranteed maximum likelihood splitting tests of a linear regression model*, Statistics **40** (2006), 465–484.

65. Allan Gut, and Josef Steinebach, *EWMA charts for detection a change-point in the drift of a stochastic process*, Sequential Analysis **23** (2004), 195–237.

66. Patsy Haccou, and Evert Mellis, *Testing for the number of change points in a sequence of exponential random variables*, Journal of Computer Simulation **30** (1988), 285–298.

67. Patsy Haccou, Evert Mellis, and Sara Van De Geer, *The likelihood ratio test for the change poin problem for exponentially distributed random variables*, Stochastic Processes and their Applications **27** (1988), 121–139.

68. Aaron L. Halpern, *Minimally selected p and other tests for a single abrupt changepoint in a binary sequence*, Biometrics **55** (1999), 1044–1050.

69. Aaron L. Halpern, *Multiple-changepoint testing for an alternating segments model of a binary sequence*, Biometrics **56** (2000), 903–908.

70. Samir Ben Hariz, and Jonathan J. Wylie, *Convergence rates for estimating a change-point with long-range dependent sequences*, Comptes Rendus Mathematique **341** (2005), 765–768.

71. Samir Ben Hariz, Jonathan J. Wylie, and Qiang Zhang, *Nonparametric change-point estimation for dependent sequences*, C.R. Acadamy of Science Paris, Series I **341** (2005). 627–630.

72. Douglas, M. Hawkins, *Testing a sequence of observations for a shift in location*, Journal of the American Statistical Association **72** (1977), 180–186.

73. Tina Herberts, and Uwe Jensen, *Optimal detection of a change point in a Poisson process for different observation schemes*, Scandinavian Journal of Statistics **31** (2004), 347–366.

74. David V. Hinkley, em Inference about the change-point in a sequence of random varialbes, Biometrika **57** (1970), 1–17.

75. David V. Hinkley, and Elizabeth A. Hinkley, *Inference about the change-point in a sequence of binomial variables*, Biometrika **56** (1970), 477–488.

76. Lajos Horváth, *The maximum likelihood method for testing changes in the parameters of normal observations*, The Annals of Statistics **21** (1993), 671–680.

77. Lajos Horváth, and Piotr Kokoszka, *Change-point detection with non-parametric regression*, Statistics **36** (2002), 9–13.

78. Lajos Horváth, and Marie Hušková, *Testing for changes using permutiations of U-statistics*, Journal of Statistical Planning and Inference **128** (2005), 351–371.

79. Lajos Horváth, and Qi-Man Shao, *Limit theorems for permutations of empirical processes with applications to change-point analysis*, Stochastic Processes and their Applications **117** (2007), 1870–1888.

80. Lajos Horváth, Mario Kühn, and Josef Steinebach, *On the performance of the fluctuation test for structural change*, Sequential Analysis **27** (2008), 126–140.

81. D. A. Hsu, *Tests for variance at an unknown time point*, Applied Statistics **26** (1977), 279–284.

82. Li Hsu, *Denoising array-based comparative genomic hybridization data using*

wavelets, Biostatistics **6** (2005), 211–226.

83. Tao Huang, Bealin Wu, Paul Lizardi, and Hongyu Zhao, *Detection of DNA copy number alterations using penalized least squares regression*, Bioinformatics **21** (2005), 3811–3817.
84. J. Huh, and K.C. Carrière, *Estimation of regression functions with a discontinuity in a derivative with local polynomial fits*, Statistics & Probability Letters **56** (2002), 329–343.
85. J. Huh, and B.U. Park, *Detection of a change point with local polynomial fits for the random design case*, Australian and New Zeland Journal of Statistics **46** (2004), 425–441.
86. Marie Hušková, *Serial rank statistics for detection of change*, Statistics & Probability Letters **61** (2003), 199–213.
87. Marie Hušková, and Claudia Kirch, *Bootstrapping confidence intervals for the change-point of time series*, Journal of Time Series Analysis **29** (2008), 947–972.
88. Marie Hušková, and Claudia Kirch, Zuzana Prášková, and Josef Steinebach, *On the detection of chnages in autoregressive time series, II. resampling procedures*, Journal of Statistical Planning and Inference **138** (2008), 1697–1721.
89. Marie Hušková, and Georg Neuhaus, *Change point analysis for censored data*, Journal of Statistical Planning and Inference **126** (2004), 207–223.
90. Stefano M. Iacus, and Nakahiro Yoshida, *Estimation for the change point of volatility in a stochastic differential equation*, Stochasic Processes and their Applications **122** (2012), 1068–1092.
91. Carla Inclán, *Detection of multiple changes of variance using posterior odds*,

Journal of Business & Economic Statistics **11** (1993), 289–300.

92. Carla Inclán, and George C. Tiao, *Use of cumulative sums of squares for retrospective detection of changes of variance*, Journal of teh American Statistical Association **89** (1994), 913–923.
93. Barry James, Kang Ling James, and David Siegmund, *Asymptotic approximations for likelihood ratio tests and confidence regions for a change-point in the mean of a multivariate normal distrtibution*, Statistica Sinica **2** (1992), 69–90.
94. Venkata K. Jandhyala, and Stergios B. Fotopoulos, *Detection and estimation of abrupt chages in the variablity of a process*, Computational Statistics & Data Analysis **40** (2002), 1–19.
95. Maarten Jansen, *Multiscale change point analysis in Poisson count data*, Chemometrics and Intelligent Laboratory Systems **85** (2007), 159–169.
96. Wei Jiang, Lianjie Shu, and Daniel W. Apley, *Adaptive CUSUM procedures with EWMA-based shift estimators*, IIE Transactions **40** (2008), 992–1003.
97. Hao Jin, Zheng Tian, and Ruibing Qin, *Subsampling tests for the mean change point with heavy-tailed innovations*, Mathematics and Computers in Simulation **79** (2009), 2157–2166.
98. Moritz Jirak, *Change-point analysis in increasing dimension*, Journal of Multivariate Analysis **111** (2012), 136–159.
99. Kees Jong, Elena Marchiori, Gerrit Meijer, A. v. d. Vaart, and Bauke Ylstra, *Breakpoint identification and smoothing of array comparative genomic hybridization data*, Bioinformatics **20** (2004), 3636–3637.
100. Durdu Karasoy, and Cem Kadilar, *A new Bayes estimate of the change point*

in the hazard function, Computational Statistics & Data Annalysis **51** (2007), 2993–3001.

101. Durdu Karasoy, and Cem Kadilar, *Modified estimators for the change point in hazard function*, Journal of Computational and Applied Mathematics bf 229 (2009), 152–157.

102. Tae-Min Kim, Yu-Chae Jung, Mun-Gan Rhyu, Myeong Ho Jung, and Yeun-Jun Chung, *GEAR: genomic enrichment analysis of regional DNA copy number changes*, Bioinformatics **24** (2008), 420–421.

103. Claudia Kirch, *Block permutation principles for the change analysis of dependent data*, Journal os Statistical Planning and Inference **137** (2007), 2453–2474.

104. Claudia Kirch, *Bootstrapping sequential change-point tests*, Sequential Analysis **27** (2008), 330–349.

105. Claudia Kirch, and Josef Steinebach, *Permutation principles for the change analysis of stochastic processes under strong invariance*, Journal of Computational and Applied Mathematics **186** (2006), 64–88.

106. Sven Knoth, *EWMA schemes with nonhomogeneous transition kernels*, Sequential Analysis **22** (2003), 241–255.

107. Asako Koike, Nao Nishida, Daiki Yamashita, and Katsushi Tokunaga, *Comparative analysis of copy number variation detection methods and database construction*, BMC Genetics **12** (2011), 29–40. http://www.biomedcentral.com/1471-2156/12/29, doi: 10.1186/1471-2156-12-29.

108. Kyungduk Ko, and Marina Vannucci, *Beyesian wavelet-based methods for the detection of multiple changes of the long memory parameter*, IEEE Transactions

on Signal Processing **54** (2006), 4461–4470.

109. Hira L. Koul, and Lianfen Qian, *Asymptotics of maximum likelihood estimator in a two-phase linear regression model*, Journal of Statistical Planning and Inference **108** (2002), 99–119.

110. P. R. Krishnaiah, Baiqi Miao, and Lincheng Zhao, *Local likelihood method in the problems related to change points*, Chinese Annals of Mathematics **11B** (1990), 363–375.

111. Kuldeep Kumar, and Berlin Wu, *Detection of change points in time series analysis with fuzzy statistics*, International Journal of Systems Science **32** (2002), 1185–1192.

112. Tze Leung Lai, and Haipeng Xing, *A simple Bayesian approach to multiple change-points*, Statistica Sinica **21** (2011), 539–569.

113. Michael Last, and Robert Shumway, *Detecting abrupt chagnes in a piecewise locally stationary time series*, Journal of Multivariate Analysis **99** (2008), 191–214.

114. Marc Lavielle, *Using penalized contrasts for the change-point problem*, Signal Processing **85** (2005), 1501–1510.

115. Sangyeol Lee, and Seongryong Na, *A nonparametric test for the change of the density function in strong mixing processes*, Statistics & Probability Letters **66** (2004), 25–34.

116. E.L. Lehmann, *Testing Statistical Hypotheses*, 2nd ed., Wiley, New York, 1986.

117. Albert M. Levin, Debashis Ghosh, Kathleen R. Cho, and Sharon L. R. Kardia, *A model-based scan statistic for identifying extreme chromosomal regions of gene*

extression in human tumors, Bioinformatics **21** (2005), 2867–2874.

118. Degui Li, and Zhengyan Lin, *A nonparametric test for the change of the density function under association*, Nonparametric Statistics **19** (2007), 1–12.

119. Youjuan Li, and Ji Zhu, *Analysis of array CGH data for cancer studies suing fused quantile regression*, Bioinformatics **23** (2007), 2470–2476.

120. Jing Lin, *A two-stage failure model for bayesian change point analysis*, IEEE Transactions on Reliability **57** (2008), 388–393.

121. Zhengyan Lin, Degui Li, and Jia Chen, *Change point estimators by local polynomial fits under a dependence assumption*, Journal of Multivariate Analysis **99** (2008), 2339–2355.

122. Jungeng Liu, E. James Harner, and Harry Yang, *Bayesian circle segmentation with application to DNA copy nubmer alteration detection*, Journal of Data Science **6** (2008), 53–73.

123. Clive R. Loader, *Change point estimation using nonparametric regression*, The Annals of Statistics **24** (1996), 1667–1678.

124. R.H. Loschi, and F.R.B. Cruz, *An analysis of the influences of some prior specifications in the identification of change points via product partition models*, Computational Statistics & Data Analysis **39** (2002), 477–501.

125. R.H. Loschi, and F.R.B. Cruz, *Applying the product partition model ot the identification of multiple change points*, Advances in Complex Systems **5** (2002), 371–387.

126. R.H. Loschi, and F.R.B. Cruz, *Extention to the product partition model: computing the probability of a change*, Computational Statistics & Data Analysis **48**

(2005), 255–268.

127. R.H. Loschi, and F.R.B. Cruz, *Bayesian identification of multiple change points in poisson data*, Advances in Complex Systems **8** (2005), 465–482.

128. Adam Al Margolin, Joel Greshock, Tara L. Naylor, Yeal Mosse, John M. Maris, Graham Bignell, Alexander L. Saeed, John Quackenbush, and Barbara L. Weber, *CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data*, Bioinformatics **21** (2005), 3308–3311.

129. Ulrich Menzefricke, *A bayesian analysis of a change in the precision of a sequence of independent normal random variables at an unknown time point*, Applied Statistics **30** (1981), 141–146.

130. Vladimir N. Minin, Karin S. Borman, Fang Fang, and Marc A. Suchard, *Dual multiple change-point model leads to more accurate recombination detection*, Bioinformatics **21** (2005), 3034–3042.

131. Hans-Georg Müller, *Change-points in nonparametric regression analysis*, The Annals of Statistics **20** (1992), 737–761.

132. Hans-Georg Müller, and Newton Wai, *Asymptotic fluctuations of mutigrams*, Statistics & Probability Letters **76** (2006), 1201–1210.

133. Chad L. Myers, Maitreya J. Dunham, S. Y. Kung, and Olga G. Troyanskaya, *Accurate detection of aneuploidies in array CGH and gene expression microarray data*, Bioinformatics **20** (2004), 3533–3543.

134. Weilin Nie, Samir Ben Hariz, Jonathan Wylie, and Qiang Zhang, *Detection of change-points of long-range dependent sequences*, C.R. Acadamy of Science Paris, Series I **347** (2009), 425–428.

135. Yoshiyuki Ninomiya, *Construction of conservative test for change-point problem in two-dimensional random fields*, Journal of Multivariate Analysis **89** (2004), 219–242.

136. Yoshiyuki Ninomiya, *Information criterion for Gaussian change-point model*, Statistics & Probability Letters **72** (2005), 237–247.

137. Yoshiyuki Ninomiya, and Atsushi Yoshimoto, *Statistical Method for detecting structural change in the growth process*, Biometrics **64** (2008), 46–53.

138. Tsan Sheng Ng, *An application of the EM algorithm to degradation modeling*, IEEE Transactions on Reliability **57** (2008), 2–13.

139. K. Nosek, and Z. Szkutnik, *A power study of k-linear-r-ahead recursive residuals test for change-point in finite sequences*, Journal of Statistical Computation and Simulation **78** (2008), 1201–1213.

140. Kyong Joo Oh, Myung Sang Moon, and Tae Yoon Kim, *Variance change point detection via artificial neural networks for data separation*, Neurocomputing **68** (2005), 239–250.

141. Adam B. Olshen, and E. S. Venkartaman, *Circular binary segmentation for the analysis of array-based DNA copy number data*, Biostatistics **5** (2004), 557–572.

142. Lena R. Olsen, Sigrunn H. Sørbye, and Fred Godtliebsen, *A scale-space approach for detecting non-stationaries in time series*, Scandinavian Journal of Statistics **35** (2008), 119–138.

143. Markus Orasch, and William Pouliot, *Tabulating weighted sup-norm functionals used in change-point analysis*, Journal of Statisical Computation & Simulation **74** (2004), 249–276.

144. E.S. Page, *Continuous Inspection Schemes*, Biometrica **41** (1954), 100–115.

145. E.S. Page, *A test for a change in a parameter occurring at an unknown point*, Biometrica **42** (1955), 523–527.

146. E.S. Page, *On problems in which a change in a parameter occurs at an unknown point*, Biometrica **44** (1957), 248–252.

147. Jianmin Pan, and Jiahua Chen, em Application of modified information criterion to multiple change point problems, Journal of Multivariate Analysis **97** (2006), 2221–2241.

148. Cheol-Woo Park, and Woo-Chul Kim, *Estimation of a regression function with a sharp change point using boundary wavelets*, Statistics & Probability Letters **66** (2004), 435–448.

149. Marcus B. Perry, Joseph J. Pignatiello, Jr, and James R. Simpson, *A magnitude-robust control chart for monitoring and estimating step changes in a Poisson rate parameter*, International Journal of Reliability, Quality and Safety Engineering **14** (2007), 1–19.

150. Marcus B. Perry, and Joseph J. Pignatiello, Jr, *Estimation of the change point of the process fraction nonconforming in SPC applications*, International Journal of Reliability, Quality and Safety Engineering **12** (2005), 95–110.

151. Marcus B. Perry, and Joseph J. Pignatiello, Jr, *A change point model for the location parameter of exponential family densities*, IIE Transactions **40** (2008), 947–956.

152. A.N. Pettitt, *A simple cumulative sum type statistic for the change-point problem with zero-one observations*, Biometrika **67** (1980), 79–84.

153. Dominique Picard, *Testing and estimating change-points in time series*, Advances in Applied Probability **17** (1985), 841–867.

154. Franck Picard, Stephane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin, *A statistical approach for array CGH data analysis*, BMC Bioinformatics **6**:27 (2005).

155. Gerlind Plonka, and Jianwei Ma, *Convergence of an iterative nonlinear scheme for denoising of piecewise constant images*, International Journal of Wavelets, Multiresolution and Information Processing **5** (2007), 975–995.

156. Odile Pons, *Estimation in a Cox regression model with a change-point at an unknown time*, Statistics **36** (2002), 101–124.

157. Richard E. Quandt, *The estimation of the parameters of a linear regression system obeying two seperate regimes*, Journal of the American Statistical Association **53** (1958), 873–880.

158. Richard E. Quandt, *Tests of the hypothesis that a linear regression system obeys two separte regimes*, Journal of the American Statistical Association **55** (1960), 324–330.

159. Fernando A. Quintana, Pilar L. Iglesias, and Heleno Bolfarine, *Bayesian identification of outliers and change-points in measurement error models*, Advances in Comples Systems **8** (2005), 433–449.

160. Alfredas Račkauskas, and Charles Suquet, *Hölder norm test statistics for epidemic change*, Journal of Statistical Planning and Inference **126** (2004), 495–520.

161. Ewaryst Rafajlowicz, and Ansgar Steland, *A binary control chart to detect small jumps*, Statistics **43** (2009), 295–311.

162. A.E. Raftery, and V.E. Akman, *Bayesian analysis of a Poisson process with a change-point*, Biometrika **73** (1986), 85–89.

163. Bonnie K. Ray, and Ruey S. Tsay, *Bayesian methods for change-point detection in long-range dependent proecesses*, Journal of Time Series Analysis **23** (2002), 687–705.

164. S. Reboul, and M. Benjelloun, *A Bayesian fusion approach to change-points analysis of processes*, IEEE: 7th International Conference on Information Fusion (FUSION) **1** (2005), 699–704.

165. S.W. Roberts, *Control chart tests based on geometric moving averages*, Technometrics **1** (1959), 239–250.

166. Yves Rozenholc, *Nonparametric tests of change-points with tapered data*, Journal of Time Series Analysis **22** (2001), 13–43.

167. Andrew L. Rukhin, *Asymptotic behavior of posterior disribution of the change-point parameter*, Journal of Statistical Planning and Inference **105** (2002), 327–345.

168. Mark R. Segal, and Joseph L. Wiemels, *Clustering of translocation breakpoints*, Journal of the American Statistical Association **97** (2002), 66–76.

169. Ashish K. Sen, and Muni S. Srivastava, *On multivariate tests for detecting change in mean*, Sankhyá **35** (1973), 173–186.

170. Ashish K. Sen, and Muni S. Srivastava, *On tests for detecting change in mean*, The Annals of Statistics **3** (1975), 98–108.

171. Ashish K. Sen, and Muni S. Srivastava, *Some one-sided tests for change in level*, Technometrics **17** (1975), 61–64.

172. Sohrab P. Shah, Xiang Xuan, Ron J. DeLeeuw, Mehrnoush Khojasteh, Wan L. Lam, Raymond Ng, and Kevin P. Murphy, *Integrating copy number polymorphisms into array CGH analysis using a robust HMM*, Bioinformatics **22** (2006), e431–e439. doi:10.1093/bioinformatics/btl238

173. Gang Shen, and Jayanta K. Ghosh, *Developing a new BIC for detecting change-points*, Journal of Statistical Planning and Inference **141** (2011), 1436–1447.

174. Xiaoping Shi, Yuehua Wu, and Baiqi Miao, *Strong convergence rate of estimators of change point and its application*, Computational Statistics and Data Analysis **53** (2009), 990–998.

175. D.O. Siegmund, and B. Yakir, *Minimax optimality of the Shiryayev-Roberts change-point detection rule*, Journal of Statistical Planning and Infernce **138** (2008), 2815–2825.

176. A.F.M. Smith, *A Bayesian approach to inference about a change-point in a sequence of random varialbes*, Biometrika **62** (1975), 407–416.

177. A.F.M. Smith, and D.G. Cook, *Straight lines with a change-point: a Bayesian analysis of some renal transplant data*, Applied Statistics, **29** (1980), 180–189.

178. A. M. Snijders, N.Nowak, R. Segraves et al., *Assembly of microarrays for genome-wide measurement of DNA copy number*, Nature Genetics, **29** (2001), 263–264.

179. G. Yu. Sofronov, *Asymptotically d-optimal test of a change-point detection*, Theory of Probability & Its Applicaitons **46** (2001), 547–548.

180. G. Yu. Sofronov, *Asymptotically d-optimal test of a posteriori cange-point detection*, Theory of Probability & Its Applications **49** (2005), 367–371.

181. Young Sook Son, and Seong W. Kim, *Bayesian single change point detection in a sequence of multivariate normal observations*, Statistics **39** (2005), 373–387.

182. M. S. Srivastava, and K. J. Worsley *Likelihood ratio tests for a change in the multivariate normal mean*, Journal of the American Statistical Association **81** (1986), 199–204.

183. D. A. Stephens, *Bayesian retrospective multiple=changepoint identification*, Applied Statistics **43** (1994), 159–178.

184. Susann Stjernqvist, Tobias Rydn, Martin Skld, and Johan Staaf, *Continuous-index hidden Markov modelling of array CGH copy number data*, Bioinformatics **23** (2007), 1006–1014.

185. Yan V. Sun, Douglas M. Jacobsen, and Sharon L. R. Kardia, *ChromoScan: a scan statistic application for identifying chromosomal regions in genomic studies*, Bioinformatics **22** (2006), 2945–2947.

186. Y. V. Sun, A.M. Levin, E. Boerwinkle, H. Robertson, and S. L. Kardia, *A scan statistic for identifying chromosomal patterns of SNP association*,Genetic Epidemiology **30** (2006), 627–635.

187. Gideon Schwarz, *Estimating the dimention of a model*, The Annals of Statistics, **6** (1978), 461–464.

188. Jun-ichi Takeuchi, and Kenji Yamanishi, *A unifying framework for detecting outliers and change points from time series*, IEEE Transactions on Knowledge and Data Engineering **18** (2006), 482–492.

189. Gua-Liang Tian, Kai Wang Ng, Kai-Can Li, and Ming Tan, *Non-iterative sampling-based Bayesian methods for identifying changepoints in the sequence*

of cases of Haemolytic uraemic syndrome, Computational Statistics and Data Analysis **53** (2009), 3314–3323.

190. J.Y. Tourneret, A. Ferrari, and A. Swami, *Cramer-Rao lower bounds for change points in additive and multiplicative noise*, Signal Processing **84** (2004), 1071–1088.

191. M.H. Vellekoop, and J.M.C. Clark, *A nonlinear filtering approach to changepoint detection problems: direct and differential-deometric methods*, SIAM Review **48** (2006), 329–356.

192. Tomáš Víšek, *The likelihood ratio method for testing changes in the parameters of double exponential observations*, Journal of Statistical Planning and Inference **113** (2003), 79–111.

193. L. Ju Vostrikova, *Detection "disorder" in multidemensional random processes*, Soviet Mathematics Doklady **24** (1981), 55–59.

194. Abraham Wald, *Note on the Consistency of the Maximum Likelihood Estimate*, The Annals of Mathematical Statistics **20** (1949), 595–601.

195. Guangming Wang, Samir Ben Hariz, Jonathan J. Wylie, and Qiang Zhang, *Change-point detection for continuous processes with high-frequency sampling*, Statistics bf 346 (2008), 467–470.

196. Lihong Wang, and Jinde Wang, *Testing and estimating for change in long memory parameter*, Journal of Statisical Computaion and Simulation **76** (2006), 317–329.

197. Lihong Wang, *Gradual changes in long memory processes with applications*, Statistics **41** (2007), 221–240.

198. Lihong Wang, *Change-in-mean problem for long memory time series models with applications*, Journal of Statistical Computation and Simulation **78** (2008), 653–668.

199. Lihong Wang, *Change-point detection with rank statistics in long-memory time-series models*, Australian and New Zeland Journal of Statistics **50** (2008), 241–256.

200. Lihong Wang, and Haiyan Cai, *Wavelet change-point estimation for long memory non-parametric random design models*, Journal of Time Series Analysis **31** (2010), 86–97.

201. Yazhen Wang, *Jump and sharp cusp detection by wavelets*, Biometrika **82** (1995), 383–397.

202. Yazhen Wang, *Change-points via wavelets for indirect data*, Statistica Sinica **9** (1999), 103–117.

203. Zhan Feng Wang, Yao Hua Wu, and Lin Cheng Zhao, *Nonparametric Inference in a simple change-point model*, Acta Mathematica Sinica, English Series **24** (2008), 1483–1496.

204. Lucas P. Watkins, and Haw Yang, *Detection of intensity change points in time-resolved single-molecule measurements*, Journal of Physical Chemistry B **109** (2005), 617–628.

205. Dean W. Wichern, Robert B. Miller, and Der-Ann Hsu, *Change in variance of first-order autoregressive time series*, Applied Statistics **25** (1976), 248–256.

206. Hanni Willenbrock, and Jane Fridlyand, *A comparison study: applying segmentation to array CGH data for downstream analyses*, Bioinformatics **21** (2005),

4084–4091.

207. M.R. Williams, and D. Kim, *Likelihood fatio tests for continuous monotone hazards with an unknown change point*, Statistics and Probability Letters **81** (2011), 1599–1603.

208. K.J. Worsley, *On the likelihood ratio test for a shift in location of normal populations*, Journal of the American Statistical Association **74** (1979), 365–367.

209. K.J. Worsley, *The power of likelihood ratio and cumulative sum tests for a change in a binomial problility*, Biometrika **70** (1983), 455–464.

210. K.J. Worsley, *Confidence regions and tests for a change-point in a squence of exponential family random variables*, Biometrika **73** (1986), 91–104.

211. C.Q. Wu, L.C. Zhao, and Y.H. Wu, *Estimation in change-point hazard function models*, Statistics & Probability Letters **63** (2003), 41–48.

212. Yanhong Wu, *Infernece for change-point and post-change mean with possible change in variance*, Sequential Analysis **24** (2005), 279–302.

213. Y. Wu, *Simultaneous change point analysis and variable selection in a regression problem*, Journal of Multivariate Analysis **99** (2008), 2154–2171.

214. Zhang Wu, Jianxin Jiao, Mei Yang, Ying Liu, and Zhaojun Wang, *An enhanced adaptive CUSUM control chart*, IIE Transactions **41** (2009), 642–653.

215. Baifang Xing, Celia M.T. Greenwood, and Shelley B. Bull, *A hierarchical clustering method for estimating copy number variation*, Biostatistics **8** (2006), 632–653.

216. Jing-Rung Yu, Gwo-Hshiung Tzeng, and Han-Lin Li, *Interval piecewise regression model with automatic change-point detection by quadratic programming*, In-

ternational Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **13** (2005), 347–361.

217. Jing-Rung Yu, and Gwo-Hshiung Tzeng, *Fuzzy multiple objective programming in an interval piecewise regression model*, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **17** (2009), 365–376.
218. S. Zacks, *Distributions of stopping times for Poisson processes with linear boundaries*, Communications in Statistics - Stochastic Models, **7** (1991), 233–242.
219. S. Zacks, *Exact determinaition of the run length destribution of a one-sided cusum procedure applied on an ordinary Poisson process*, Sequential Analysis **23** (2004), 159–178.
220. Nancy R. Zhang, and David O. Siegmund, *A modified bayes information criterion with applications to the analysis of camparative genomic hybridization data*, Biometrics **63** (2007), 22–32.

VITA

Paul Jacob Plummer was born on July 18, 1965, in Canoga Park, California. In 1983 he graduated from Granada Hills High School in Granada Hills, California. He received his B.S.E in Mathematics and Social Studies in 1990, and his M.S. in Mathematics in 1995, from the University of Central Missouri, Warrensburg, Missouri. He has taught at both the secondary and higher education level, and upon graduation will continue his career in education as an Assistant Professor at the University of Central Missouri.