1-1-2015

Change Point Detection and Estimation in Sequences of Dependent Random Variables

Benjamin Cortese
Syracuse University

ABSTRACT

Two change point detection and estimation procedures for sequences of dependent binary random variables are proposed and their asymptotic properties are explored. The two procedures are a dependent cumulative sum statistic (DCUSUM) and a dependent likelihood ratio test (LRT) statistic, which are generalizations of the independent CUSUM and LRT statistics.

A one step Markov dependence is assumed between consecutive variables in the sequence, and the performance of the DCUSUM and dependent LRT are shown to have substantially better size and power performance than their independent counterparts. In most cases, a comparison of the dependent procedures via simulation shows that the dependent LRT provides a more powerful test, while the DCUSUM test has better size performance.

The asymptotic distribution of the DCUSUM test is found to be a weighted sum of squared Brownian bridge processes and an approximation to calculate p-values is discussed. A Worsley type upper bound for p-values is provided as an alternative. The asymptotic distribution of the dependent LRT is unknown, but the tail probabilities are found to be empirically bounded by a χ^2_6 and a χ^2_7 random variable through a simulation study. A bootstrap algorithm to estimate p-values for the dependent LRT is discussed.

Extensions of these procedures to multiple sequences and multinomial random variables are discussed, and a new statistic, the maximal change count statistic, is proposed. An application of the multiple sequence procedures to clustered time series models is provided. The asymptotic properties of the generalized procedures are reserved for future research.

Change Point Detection and Estimation in Sequences of Dependent Random Variables

by

Benjamin David Cortese

B.S., The College at Brockport: State University of New York - Brockport, 2009

M.A., Syracuse University - Syracuse, 2011

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mathematics.

Syracuse University

August 2015

Acknowledgments

I would like to thank my advisor, Dr. Hyune-Ju Kim, for being my mentor and guide during my graduate career. I am indebted to my colleagues at Syracuse University for providing help through my years of graduate study. I am especially grateful to Dr. Thomas John, Dr. Steve Stehman, Dr. William Volterman, Dr. Leonid Kovalev, Dr. Ted Cox, Dr. Ken Foster, and to Dr. Pinyuen Chen for encouraging me to pursue statistics early in my graduate career. Additionally, I would like to thank my friends and family, especially my parents, Steve and Deb Cortese, for supporting me through this journey.

Finally, I would like to dedicate this dissertation to my wife Lynn - thank you for your love and support through all of the long days and nights.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The change point problem has been studied for decades and recently has experienced an increase in popularity. These problems arise in a variety of situations including regression, clustering, sequences of random variables, control charts, etc.

The aim of this dissertation is to detect and estimate a change point τ in a sequence of dependent random variables $y = \{x_1, x_2, \ldots, x_n\}$ of length n. Several methods are proposed to detect the existence of a change point $1 < \tau < n$ for the sequence y. If such a change occurred, the location of the change, τ, will be estimated.

Early tests to determine if two random variables are stochastically different include the Mann-Whitney U test [25] and the Kolmogorov-Smirnov test. These can be used to test if a change is significant by segmenting the data at the suspected change point and comparing the two approximate distributions from each segment. This dissertation aims to extend existing tests by relaxing some of the assumptions.

The first method to tackle the change point problem was formally stated by Page [28, 29] and much of the work post 1955 have referenced these papers. After Page, Hinkley [13] discussed the maximum likelihood estimates (MLE) and likelihood ratio test (LRT) for a change in parameter θ, and derived the asymptotic distribution for *iid* sequences with general *pdf* $f(x, \theta)$ and the *iid* normal case.

Much of the early work in change point detection assumed the random variables in the sequence were continuous. For the purposes of this dissertation, the distributions are restricted to discrete random variables. Several methods have been explored to tackle discrete sequences including MLE and LRT methods, cumulative sum statistics (CUSUM) and maximally selected χ^2 statistics.

The main results of this dissertation focus on sequences of Bernoulli random variables. Pettitt [31] introduced the CUSUM method for discrete random variables taking on the values of 0 or 1. Miller and Siegmund [26] proposed the maximally selected χ^2 method to detect a change in a sequence by selecting cut points and comparing the two sequences. Halpern [10] compared the performance of several methods for binary random variables and noted that there is no uniformly most powerful test.

The obvious extension from Bernoulli random variables is to explore the location of change points in a sequence of binomial random variables. Hinkely and Hinkley [14] extend the results from an earlier paper, Hinkley [13], to binomial sequences and show that the likelihood ratio test corresponds to the distribution of a random walk. A power comparison of the likelihood ratio test to the CUSUM test is performed by Worsley [41], who shows that the LRT is slightly less powerful than CUSUM in the center of the sequence, and the opposite is true at the tails.

The extension from binomial to multinomial random variables is quite natural and has been studied extensively over the past two decades. Three of the most common statistics for testing for a change in a multinomial sequence are the likelihood ratio, CUSUM and maximal χ^2 statistics. Horváth and Serbinowska [17] describe the asymptotic distribution of the LR statistics. MacNeill [22] first studied CUSUM statistics and Robbins et al. [34] clearly summarized the CUSUM and maximal χ^2 results. Robbins et al. also provided an approximation for tail probabilities for the asymptotic distribution of these three statistics.

All of the change point detection and estimation methods mentioned above rely on the assumption that any two variables x_i and x_j where $i \neq j$ in the sequence $\{x_i\}$ are inde-

pendent. In many situations, the variables in the sequence may exhibit some dependence structure. The literature on change point detection and estimation for dependent sequences is limited. Most of the work done in this area is by Krauth [20, 21]. One contribution of this dissertation is to extend the results for independent sequences to dependent sequences, assuming a one step Markov dependence on the variables x_i.

This dissertation is organized as follows. The remainder of this chapter is a review of the change point detection and estimation procedures for independent sequences of discrete random variables, followed by the assumptions and hypotheses in the one step Markov dependence case. Chapter two is an extension of the CUSUM statistic, and Chapter three is an extension of the likelihood ratio statistic. Chapter four is comprised of simulations to compare the level and power of all methods for different parameter values. Chapter five describes extensions to multinomial and multiple sequence methods and provides the main motivating application of this work.

1.1 Maximum Likelihood Estimation of Change Point Locations in Independent Sequences of Random Variables

This section provides a review of the change point estimation procedures for independent sequences of discrete random variables. In particular, the maximum likelihood estimation for Bernoulli sequences is discussed under certain assumptions. Other methods include weighted squares [4, 39] and single-switch multinomial logistic models [33], but those details are omitted.

Let $y_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ for $i = 1, 2, \ldots, s$ represent s sequences of random variables, each of length n. There are two types of procedures when working with multiple sequences of random variables. The first procedure is called a *single path* procedure. A single path

procedure is restricted to the information from a single sequence y_i and omits the other sequences. A *multi path* procedure uses information from all of the s sequences $\{y_i\}_{i=1}^{s}$. While multi path procedures use significantly more data in parameter estimation, these methods tend to have more restrictive assumptions. The main results of this dissertation are focused on single path procedures for the sequence $y = \{x_1, x_2, \ldots, x_n\}$. Generalizations to multi path procedures are briefly discussed in Chapter 5.

Before discussing the detection techniques, it is first assumed that a change point $1 < \tau < n$ exists in the sequence y. The procedure to estimate the location of this change using the maximum likelihood estimate (MLE) is described below.

Under the distributional assumptions of this dissertation, each x_j in the sequence y is assumed to be a Bernoulli(p) random variable. For each possible change point $1 < t < n$, the null model is:

$$x_j \sim \text{Bernoulli}(p) \text{ for } 1 \leq j \leq n,$$

and the alternative model is:

$$x_j \sim \text{Bernoulli}(p(1)) \text{ for } j \leq t \text{ and } x_j \sim \text{Bernoulli}(p(2)) \text{ for } j > t.$$

The log likelihood function for each x_j is the log likelihood function of a Bernoulli random variable, that is:

$$\log f(x_j \mid p) = \log(p^{x_j}(1-p)^{1-x_j}) = x_j \log\left(\frac{p}{1-p}\right) + \log(1-p). \tag{1.1}$$

For a fixed time t, the maximum likelihood estimates of the proportions p, $p(1)$ and $p(2)$ are given by $\hat{p}$, $\hat{p}(1)$ and $\hat{p}(2)$ below:

$$\hat{p} = \frac{\sum_{j=1}^{n} x_j}{n}, \quad \hat{p}(1) = \frac{\sum_{j=1}^{t} x_j}{t} \quad \text{and} \quad \hat{p}(2) = \frac{\sum_{j=t+1}^{n} x_j}{n-t}. \tag{1.2}$$

The corresponding log likelihood function for a change point in the sequence y at time t is:

$$\begin{aligned}\log L(t|y) &= \log\left(\prod_{j=1}^{t} f(x_j \mid \hat{p}(1)) \prod_{j=t+1}^{n} f(x_j \mid \hat{p}(2))\right) \\ &= t\log(1-\hat{p}(1)) + (n-t)\log(1-\hat{p}(2)) \\ &\quad + \log\left(\frac{\hat{p}(1)}{1-\hat{p}(1)}\right)\sum_{j=1}^{t} x_j + \log\left(\frac{\hat{p}(2)}{1-\hat{p}(2)}\right)\sum_{j=t+1}^{n} x_j. \end{aligned} \tag{1.3}$$

This function takes on a finite number of values, so maximization via differentiation is not possible. Instead, the estimate of the change point location $\hat{\tau}$ for the sequence y is found via grid search. The estimate is given below:

$$\begin{aligned}\hat{\tau} = \underset{1<t<n}{\arg\max} \Bigg\{ & t\log(1-\hat{p}(1)) + (n-t)\log(1-\hat{p}(2)) \\ & + \log\left(\frac{\hat{p}(1)}{1-\hat{p}(1)}\right)\sum_{j=1}^{t} x_j + \log\left(\frac{\hat{p}(2)}{1-\hat{p}(2)}\right)\sum_{j=t+1}^{n} x_j \Bigg\}. \end{aligned}$$

1.2 Change Point Detection in Independent Sequences of Random Variables

This section provides a review of change point detection procedures for independent sequences of discrete random variables. In particular, CUSUM statistics, maximal χ^2 statistics, and likelihood ratio tests for Bernoulli sequences are discussed under certain assumptions.

The details of the model and assumptions for this section can be found in Section 1.1. Keep in mind that under H_0, detection techniques no longer assume that a change point τ exists.

1.2.1 CUSUM

There is a wide class of situations where cumulative sum (CUSUM) statistics are used to detect changes in sequences. The uses include, but are not limited to, change point problems in sequences of random variables, regression, and control charts.

The CUSUM statistic discussed below is a weighted sum of the random variables in the sequence y for each fixed time $1 < t < n$ and is used to to determine if the proportion of events before and after t are statistically different. The specific events for this CUSUM statistic are $\{x_j = 1\}$. An excellent summary of the CUSUM statistic can be found in Robbins et al. [34] and the references therein.

The hypotheses for a test using the CUSUM statistic are:

$$H_0 : x_j \text{ are } iid \text{ for all times } 1 \leq j \leq n,$$
$$H_a : \mathrm{E}(x_j) \text{ shifts at some time } \tau, 1 < \tau < n.$$

These hypotheses test for a mean shift. The random variables x_j follow a Bernoulli(p) distribution, so the alternative hypothesis is equivalent to:

$$H_a : \text{There is a value } \tau \text{ such that } p = p(1) \text{ for } 1 \leq j \leq \tau$$
$$\text{and } p = p(2) \text{ for } \tau + 1 \leq j \leq n.$$

Define the weighted sum S_t for a fixed time t as:

$$S_t = \sum_{j=1}^{t} x_j - \frac{t}{n} \sum_{j=1}^{n} x_j = \sum_{j=1}^{n} a_j x_j, \quad \text{where} \quad a_j = \begin{cases} 1 - \frac{t}{n} & \text{if } 1 \leq j \leq t, \\ -\frac{t}{n} & \text{if } t+1 \leq j \leq n. \end{cases}$$

The corresponding CUSUM statistic for sequence y at time t is $\mathrm{CUSUM}_t = S_t/\sqrt{n}$. It is

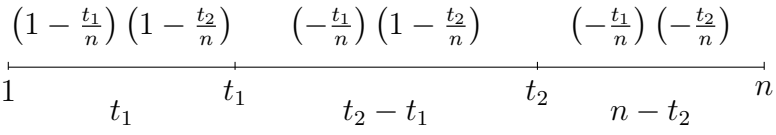
clear that $\mathrm{E}(\mathrm{CUSUM}_t) = \mathrm{E}(S_t/\sqrt{n}) = 0$. The variance calculation is given below:

$$\begin{aligned}
\mathrm{Var}(\mathrm{CUSUM}_t) &= \mathrm{Var}(S_t/\sqrt{n}) \\
&= \mathrm{Var}\left(\sum_{j=1}^{n} \frac{a_j}{\sqrt{n}} x_j\right) \\
&= \frac{1}{n}\sum_{j=1}^{n} a_j^2 \mathrm{Var}(x_j) \\
&= \frac{1}{n}\left[t\left(1-\frac{t}{n}\right)^2 + (n-t)\left(-\frac{t}{n}\right)^2\right] p(1-p) \\
&= p(1-p)\frac{t}{n}\left(1-\frac{t}{n}\right) \\
&= \sigma_t^2. \qquad (1.4)
\end{aligned}$$

Under H_0, the expected value of CUSUM_t is zero and the variance is $p(1-p)(t/n)(1-t/n)$. The CUSUM_t statistic can be viewed as the number of times $\{x_j = 1\}$ occurred up to time t compared to the total number of times $\{x_j = 1\}$ occurred in the sequence, scaled for the difference in lengths of segments. It is well known that this method has trouble detecting mean shifts at the edges of the sequence because of the nonuniform variance.

Following Robbins et al. [34], define $T_t = \mathrm{CUSUM}_t/\hat{\sigma}_t = S_t/\hat{\sigma}_t\sqrt{n}$, where $\hat{\sigma}_t$ is any consistent estimator of σ_t. For the purposes of this dissertation, define $\hat{\sigma}_t = \hat{p}(1-\hat{p})\frac{t}{n}\left(1-\frac{t}{n}\right)$, where $\hat{p}$ is the MLE defined in (1.2). Fix two values $0 < l < h < 1$ and let t_1 and t_2 be any two fixed times satisfying $nl < t_1 < t_2 < nh$. For discussion on the choice of l and h, see Miller and Siegmund [26]. First, the covariance of the CUSUM statistic for two times $t_1 < t_2$ is calculated.

The coefficients for the sum in equation (1.5) are counted using the number line below. The value of the coefficient is above the number line with the count below:



For notational purposes, define b_j to be the first value in the product above the number line for each section and c_j to be the second value.

$$\begin{aligned}
\text{Cov}(\text{CUSUM}_{t_1}, \text{CUSUM}_{t_2}) &= \frac{1}{n}\text{Cov}(S_{t_1}, S_{t_2}) \\
&= \frac{1}{n}\sum_{j=1}^{n} b_j c_j \text{Var}(x_j) \\
&= p(1-p)\left[\frac{t_1}{n}\left(1-\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)\right. \\
&\qquad + \left(\frac{t_2}{n}-\frac{t_1}{n}\right)\left(-\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right) \\
&\qquad \left. + \left(1-\frac{t_2}{n}\right)\left(-\frac{t_1}{n}\right)\left(-\frac{t_2}{n}\right)\right] \\
&= p(1-p)\frac{t_1}{n}\left(1-\frac{t_2}{n}\right). \qquad (1.5)
\end{aligned}$$

The MLE $\hat{p}$ is a consistent estimator of p, hence the covariance of T_{t_1} and T_{t_2} is:

$$\begin{aligned}
\text{Cov}(T_{t_1}, T_{t_2}) &= \frac{\text{Cov}(\text{CUSUM}_{t_1}, \text{CUSUM}_{t_2})}{\hat{\sigma}_{t_1}\hat{\sigma}_{t_2}} \\
&= \frac{p(1-p)\frac{t_1}{n}\left(1-\frac{t_2}{n}\right)}{\sqrt{\left[\hat{p}(1-\hat{p})\frac{t_1}{n}\left(1-\frac{t_1}{n}\right)\right]\left[\hat{p}(1-\hat{p})\frac{t_2}{n}\left(1-\frac{t_2}{n}\right)\right]}} \\
&\to \frac{p(1-p)\eta_1(1-\eta_2)}{p(1-p)\sqrt{\eta_1(1-\eta_1)\eta_2(1-\eta_2)}} \\
&= \left(\frac{\eta_1(1-\eta_2)}{(1-\eta_1)\eta_2}\right)^{1/2}, \qquad (1.6)
\end{aligned}$$

where $\lim_{n\to\infty} t_1/n = \eta_1$ and $\lim_{n\to\infty} t_2/n = \eta_2$.

Theorem 1 of Robbins et al. [34] states that, for fixed bounds $0 < l < h < 1$, under the null hypothesis that all x_t are *iid* for all t:

$$T^2_{\max} = \max_{l \le t/n \le h} T_t^2 \xrightarrow{\mathcal{D}} \sup_{l \le \eta \le h} \frac{B^2(\eta)}{\eta(1-\eta)}, \qquad (1.7)$$

where $B(\eta)$ is a Brownian bridge process on the interval $[0, 1]$.

1.2.2 Maximal χ^2 Statistics

The maximal χ^2 statistic is another approach to change point detection in sequences of random variables. Similar to CUSUM statistics, maximal χ^2 statistics use the number of times the event $\{x_j = 1\}$ occurred up to a fixed time t. The major difference is that comparisons are made to expected frequencies that are calculated from the MLEs of $p(1)$ and $p(2)$. The asymptotic distributions of these statistics coincides with that of the likelihood ratio statistic as discussed in Section 1.2.3.

The notation for maximal χ^2 statistics is slightly different than in Section 1.2.1. Let $n_{j,k} = 1_{\{x_{j,k}=k\}}$ for $k = 0, 1, \ldots, K$, where K is the number of possible outcomes for each variable in the sequence minus one. If $x_{j,k}$ are Bernoulli random variables then $K = 1$.

Fix a time t and define $O_{t,k} = \sum_{j=1}^{t} n_{j,k}$, $O^*_{t,k} = \sum_{j=t+1}^{n} n_{j,k}$ and $O_k = O_{n,k} = \sum_{j=1}^{n} x_{j,k}$. With these definitions in mind, one can think of $O_{t,k}$ as the number of times the sequence y takes on the value k over the first t time points, $O^*_{t,k}$ the number of times the sequence y takes on the value k over the last $n-t$ time points, and O_k as the total amount of time spent equal to k. Let $\boldsymbol{p} = (p_0, p_1)$ represent the vector of probabilities for the Bernoulli trials for any time j. The hypotheses to test are:

$$
\begin{aligned}
&H_0 : p_{j,k} = p_k \text{ for all } j, k, \\
&H_a : \text{There is a change point } \tau \text{ such that } p_{j,k} = p_{k,1},\ j \le \tau \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{and } p^*_{j,k} = p_{k,2},\ j > \tau.
\end{aligned}
$$

Define the expected counts to be the values of the MLEs described in equation (1.2). That is, $\widehat{\mathrm{E}(O_{t,k})} = t\hat{p}_k = tO_k/n$ and $\widehat{\mathrm{E}(O^*_{t,k})} = (n-t)\hat{p}_k = (n-t)O_k/n$, then the test statistic for a change point at time t is defined as in Robbins et al. [34]:

$$
\chi^2_t = \sum_{k=0}^{K} \left(\frac{\left(O_{t,k} - \widehat{\mathrm{E}(O_{t,k})}\right)^2}{\widehat{\mathrm{E}(O_{t,k})}} + \frac{\left(O^*_{t,k} - \widehat{\mathrm{E}(O^*_{t,k})}\right)^2}{\widehat{\mathrm{E}(O^*_{t,k})}} \right). \tag{1.8}
$$

Define $B^{(d)}(\eta) = \sum_{m=1}^{d} B_m^2(\eta)$ to be the sum of d independent Brownian bridge processes. Similar to CUSUM$_t$, Robbins et al. [34] shows that under H_0:

$$\chi^2_{\max} = \max_{l \le t/n \le h} \chi_t^2 \xrightarrow{\mathcal{D}} \sup_{l \le \eta \le h} \frac{B^{(K)}(\eta)}{\eta(1-\eta)} = \sup_{l \le \eta \le h} \frac{B^2(\eta)}{\eta(1-\eta)}. \tag{1.9}$$

Horvath and Serbinowska [17] consider a weighted maximal χ^2 statistic related to the Kolmogorov-Smirnov statistic to account for the slow convergence of the maximal χ^2 statistic. They show a similar asymptotic result to Robbins et al. Define:

$$Z_{n_4} = \max_{1 \le t \le n} \frac{O_{t,k} O^*_{t,k}}{{O_k}^2} \chi_t^2, \text{ where } \chi_t^2 \text{ is defined in equation (1.8) :} \tag{1.10}$$

then by Theorem 1.2 in [17]:

$$Z_{n_4} \xrightarrow{\mathcal{D}} \sup_{0 \le t \le 1} B^{(K)}(t) = \sup_{0 \le \eta \le 1} B^2(\eta). \tag{1.11}$$

1.2.3 Likelihood Ratio Test

The likelihood ratio to detect a change in a multinomial sequence is defined as the ratio of the null and alternative likelihood functions. The hypotheses are:

$$\begin{aligned} H_0 &: x_j \sim \text{Bernoulli}(p) \text{ for all } 1 \le j \le n, \\ H_a &: \text{There is } 1 < \tau < n \text{ such that } x_j \sim \text{Bernoulli}(p(1)) \text{ for } 1 \le j \le \tau \\ &\qquad \text{and } x_j \sim \text{Bernoulli}(p(2)) \text{ for } \tau + 1 \le j \le n \\ &\qquad \text{where } p(1) \ne p(2). \end{aligned}$$

Let p, $p(1)$ and $p(2)$ represent the unknown Bernoulli parameter values for variables x_j of sequence y before t and after t respectively, for any $1 \le t \le n$. Then $\hat{p}$, $\hat{p}(1)$, and $\hat{p}(2)$ can be thought of as the sample proportion of times the event $x_j = 1$ occurred before t and after t respectively, for any time $1 \le t \le n$. The MLEs of these values are given by equation

(1.2).

Under H_0, there is no change in the parameter p, so the log likelihood function for the sequence y is defined as:

$$\log(L(t|y)_{H_0}) = \sum_{1 \le j \le n} x_j \log \hat{p}.$$

The expressions for the alternative log likelihood function are given by equation (1.3). Notationally, Horvath and Serbinowska [17] define the likelihood ratio at a fixed time t as:

$$\Lambda_t = \frac{L(t|y)_{H_0}}{L(t|y)_{H_a}} \quad \text{and} \quad Z_{t,1} = \max_{1<t<n}(-2\log\Lambda_t). \tag{1.12}$$

Because $-2\log\Lambda_t$ is asymptotically a χ^2 random variable, the likelihood ratio statistic $Z_{t,1}$ has the same asymptotic behavior as the maximally selected χ^2 method. This result is summarized by Theorems 1.1 and 1.2 in Horvath and Serbinowska [17]. The weighted and unweighted asymptotic distributions were discussed in Section 1.2.2.

1.3 Tail Probability Approximations for Change Point Detection

The change point detection techniques of Section 1.2 require a method to calculate p-values. When the null distribution is known, exact or approximate p-values may be calculated in the usual way. If the null distribution is unknown or too complex to approximate, a Worsley type upper bound may provide a rough upper bound of the p-value. Both methods are discussed in this section.

1.3.1 Tail Approximation for $\sup_{l\leq\eta\leq h} \frac{B^2(\eta)}{\eta(1-\eta)}$

The general result from Robbins et al. [34] is restated below and applied to resulting asymptotic distributions of the CUSUM, maximal χ^2, and likelihood ratio statistics for the independent case. Recall that $B^{(d)}(\eta) = \sum_{m=1}^{d} B_m^2(\eta)$ where $B_m(\eta)$ are Brownian bridge processes on $[0,1]$.

$$\Pr\left(\sup_{l\leq\eta\leq h} \frac{B^{(d)}(\eta)}{\eta(1-\eta)} \geq x\right) = \frac{x^{d/2}e^{-x/2}}{2^{d/2}\Gamma(d/2)} \times \left[\left(1-\frac{d}{x}\right)\log\left(\frac{(1-l)h}{l(1-h)}\right) + \frac{4}{x} + O\left(\frac{1}{x^2}\right)\right]. \tag{1.13}$$

In equation (1.13), $O(1/x^2)$ denotes a remainder that tends to zero as $x \to \infty$ at least as fast as $1/x^2$ and $\Gamma(\cdot)$ denotes the standard gamma function.

It is clear from the expressions (1.7) and (1.9) that both CUSUM and maximal χ^2 statistics have the same asymptotic distribution. The discussion at the end of Section 1.2.3 indicates that the likelihood ratio statistic has the same asymptotic distribution as the maximal χ^2 statistic. Therefore, all three of these statistics have equivalent asymptotic distributions. For appropriately chosen values $0 < l < h < 1$, the common distribution is:

$$\sup_{l\leq\eta\leq h} \frac{B^{(1)}(\eta)}{\eta(1-\eta)} = \sup_{l\leq\eta\leq h} \frac{B^2(\eta)}{\eta(1-\eta)}.$$

The p-value approximation for an observed test statistic value T from any one of these three methods, assuming $x_j \sim \text{Bernoulli}(p)$ under H_0, is given by:

$$\Pr\left(\sup_{l\leq\eta\leq h} \frac{B^2(\eta)}{\eta(1-\eta)} \geq T\right) \approx \left(\frac{Te^{-T}}{2\pi}\right)^{1/2} \times \left[\left(1-\frac{1}{T}\right)\log\left(\frac{(1-l)h}{l(1-h)}\right) + \frac{4}{T} + O\left(\frac{1}{T^2}\right)\right]. \tag{1.14}$$

This approximation will be used to evaluate the performance of these statistics as well

as to make comparisons to the proposed methods when the independence assumption is generalized.

1.3.2 Worsley Type Upper Bound

When the independence assumption is removed, the test statistic of interest is the maximum of correlated variables $T_{\max} = \max\{T_t\}_{t=1}^{n}$. A crude upper bound for the tail probability $\Pr(T_{\max} > T)$ is given by the Bonferroni inequality:

$$\Pr(T_{\max} > T) = \Pr\left(\bigcup_{t=1}^{n} T_t > T\right) \leq \sum_{t=1}^{n} \Pr(T_t > T).$$

An improvement on this is made by Worsley [40]. Theorem 1 of his paper accounts for the correlation between all events $\{T_{t_1} > T\}$ and $\{T_{t_2} > T\}$ and is restated below in the context of Bernoulli trials with m-dependence, which will be defined in Section 1.4.2.

$$\Pr(T_{\max} > T) \leq \sum_{t=1}^{n} \Pr(T_t > T) - \sum_{|t_2 - t_1| \leq m} \Pr\left(\{T_{t_1} > T\} \cap \{T_{t_2} > T\}\right). \tag{1.15}$$

When the distribution of T_t, and joint distribution of statistics T_{t_1}, T_{t_2} are known, equation (1.15) provides an alternative method to approximate an upper bound for the p-value in a change point detection problem.

Often the exact covariance structure between all pairs of events T_{t_1} and T_{t_2} is difficult to obtain. If an incorrect structure is specified, equation (1.15) can lead to negative upper bounds on the p-value of $\Pr(T_{\max} > T)$. A simpler and often more appropriate bound, stated in Worsley [40] as Corollary 1, only requires the covariance between consecutive times t_1 and $t_2 = t_1 + 1$. The alternative upper bound is restated below:

$$\Pr(T_{\max} > T) \leq \sum_{t=1}^{n} \Pr(T_t > T) - \sum_{t=1}^{n-1} \Pr\left(\{T_t > T\} \cap \{T_{t+1} > T\}\right). \tag{1.16}$$

1.4 Assumptions and Hypotheses for Dependent Sequences

The change point detection and estimation techniques in Sections 1.1 and 1.2 rely on the assumption that the elements of the sequence $y = \{x_1, \ldots x_n\}$ are independent. In many situations, this assumption is violated. There are various possible dependence structures that may be assumed on the variables x_t.

One major motivation of the results in this dissertation is to detect changes in a clustering scheme, where x_t denote cluster membership values for sequence y at time t. This model assumes that a random variable x_t is more likely to remain at the same value from time t to time $t+1$, unless a change occurs. Details of this application are discussed in Section 5.4.

A natural representation of this is to assume a *one-step Markov dependence* between consecutive variables. For a specific variable x_t at time t, this type of dependence structure gives information about the next variable in the sequence x_{t+1} by defining a matrix of transition probabilities.

It is assumed that x_t follows a Bernoulli(p) distribution. The one-step Markov dependence assumption adds a transition matrix to the structure of each sequence. Define a *state* of the transition matrix to be a possible value that the random variable x_t can achieve at any time t. Notice that the states are the same for all t. Let u and v be two states. If it is known that the random variable $x_t = u$ and $x_{t+1} = v$, then the transition probability from state u at time t to v at time $t+1$ is given by:

$$P_{u,v,t,t+1} = \Pr(x_{t+1} = v \mid x_t = u). \tag{1.17}$$

For notational purposes, a single subscript t is used to denote the transition probability from time t to time $t+1$, that is, $P_{u,v,t,t+1} = P_{u,v,t}$. If the transition probabilities are the same for all time points t, $P_{u,v,t}$ is denoted as P_{uv}.

1.4.1 Hypotheses

Under the null hypothesis of no change, the variables x_t of the sequence y are assumed to follow a Bernoulli(p) distribution with one step Markov dependence defined by the transition probabilities P_{uv}. That is, each x_t has the same parameter value p and transition probability P_{uv}, independent of the time t.

Under the alternative hypothesis of an abrupt change at an unknown time τ, it is assumed that the Bernoulli parameter p and transition probabilities P_{uv} are disrupted at the time of the change. Specifically, the membership values $\{x_t\}_{t=1}^{\tau}$ are assumed to be independent of the values $\{x_t\}_{t=\tau+1}^{n}$. The formal hypotheses are stated below. Another possible alternative hypothesis is discussed at the end of this section.

H_0 : $x_t \sim$ Bernoulli(p) with transition probabilities P_{uv} for all times t,

H_a : There exists τ, $1 < \tau < n$, such that

$x_t \sim$ Bernoulli($p(1)$) for all $1 < t \leq \tau$ and $x_t \sim$ Bernoulli($p(2)$) for all $\tau < t \leq n$

where $p(1) \neq p(2)$ and the events after the change are independent of the events prior to the change.

Restricting to the Bernoulli case, there are exactly two states $u = 0$ or 1. The transition matrix given below is read from the state on the left at time t to the state on top at time $t+1$. This construction forces the values in the rows of the transition matrix to sum to one. Define the transition matrix for sequence y from time t to $t+1$ as $\boldsymbol{P}_t$, then under H_0 the parameters of x_t for any t are:

$$p \text{ and } \boldsymbol{P} = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}, \tag{1.18}$$

and under H_a, the parameters of x_t before τ and after τ are:

$$p(1) \text{ and } \boldsymbol{P}(1) = \begin{pmatrix} P_{00}(1) & P_{01}(1) \\ P_{10}(1) & P_{11}(1) \end{pmatrix} \text{ for } 1 \leq t \leq \tau,$$
$$p(2) \text{ and } \boldsymbol{P}(2) = \begin{pmatrix} P_{00}(2) & P_{01}(2) \\ P_{10}(2) & P_{11}(2) \end{pmatrix} \text{ for } \tau < t \leq n. \tag{1.19}$$

Under H_0, these assumptions lead to a solvable system of equations. These follow from the law of total probability and other elementary probability rules.

$$\begin{aligned} \Pr(x_t = 0) &= \Pr(x_t = 0 | x_{t-1} = 0)\Pr(x_{t-1} = 0) \\ &\quad + \Pr(x_t = 0 \mid x_{t-1} = 1)\Pr(x_{t-1} = 1), \\ \Pr(x_t = 1) &= \Pr(x_t = 1 \mid x_{t-1} = 1)\Pr(x_{t-1} = 1) \\ &\quad + \Pr(x_t = 1 \mid x_{t-1} = 0)\Pr(x_{t-1} = 0), \\ 1 &= \Pr(x_t = 0 \mid x_{t-1} = 0) + \Pr(x_t = 1 \mid x_{t-1} = 0), \\ 1 &= \Pr(x_t = 0 \mid x_{t-1} = 1) + \Pr(x_t = 1 \mid x_{t-1} = 1). \end{aligned} \tag{1.20}$$

The equations (1.20) can be written using (1.18) as:

$$\begin{aligned} (1 - p) &= P_{00}(1 - p) + P_{10}p, \\ p &= P_{11}p + P_{01}(1 - p), \\ 1 &= P_{00} + P_{01}, \\ 1 &= P_{10} + P_{11}. \end{aligned}$$

Solving the system gives the following solution in terms of p with one free variable P_{11}:

$$P_{10} = 1 - P_{11},$$

$$P_{01} = (1 - P_{11})p/(1-p),$$
$$P_{00} = (1 - 2p + P_{11}p)/(1-p). \tag{1.21}$$

Notice that each of the transition probabilities must satisfy $0 < P_{uv} < 1$. This leads to a boundary restriction on the values of P_{11} given in the right hand side of equation (1.22):

$$0 < P_{00} < 1 \Leftrightarrow 0 < (1 - 2p + P_{11}p)/(1-p) < 1 \Leftrightarrow 2 - \frac{1}{p} < P_{11} < 1. \tag{1.22}$$

It may seem natural to add additional restrictions to the values of $\boldsymbol{P}$, depending on the data that is being modeled. One such restriction is to assume that the transition probabilities from one cluster to the other are the same, that is, $P_{01} = P_{10}$. Another intuitive assumption is that the probabilities of remaining in the same cluster are equal, that is, $P_{00} = P_{11}$. Unfortunately, the one step Markov dependence assumption does not allow for either of these restrictions.

Proposition 1.4.1 *Suppose* $p \in (0,1) \setminus \{\frac{1}{2}\}$. *If* $P_{00} = P_{11} := P_0$ *or* $P_{01} = P_{10} := P_1$ *then* $P_0 = 1$ *and* $P_1 = 0$.

Proof Suppose $P_{00} = P_{11} := P_0$. Substitution into (1.21) yields:

$$\begin{aligned} P_{00} &= (1 - 2p + P_{11}p)/(1-p), \\ P_0 &= (1 - 2p + P_0 p)/(1-p), \\ (1-2p)P_0 &= 1 - 2p, \\ P_0 &= 1. \end{aligned}$$

Next, suppose $P_{01} = P_{10} := P_1$. Substitution into (1.21) again gives $P_{11} = 1 - P_1$ and:

$$\begin{aligned} P_{01} &= (1 - P_{11})p/(1-p), \\ P_1 &= (1 - (1 - P_1))p/(1-p), \end{aligned}$$

$$P_1(1-2p) = 0,$$
$$P_1 = 0.$$

In both cases, $P_{10} + P_{11} = 1$ and $P_{01} + P_{00} = 1$. These two equations conclude the proof. ∎

The *stationary distribution* $\boldsymbol{\pi}$ of a Markov chain is formally defined as a vector whose entries are non-negative, sum to one, and satisfy $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$. A Markov chain is *irreducible* if it is possible to get to any state u from any state v. A Markov chain is *aperiodic* if the return to all states can occur at irregular times. By Theorem 6.6.4 in Durrett [8], if a Markov chain is irreducible and aperiodic with stationary distribution $\boldsymbol{\pi}$, then $\lim_{w\to\infty} \boldsymbol{P}^w = \boldsymbol{\pi}$. The distribution is easily calculated using the reparametrization of the components of $\boldsymbol{P}$ above.

Lemma 1.4.2 *As* $|t_1 - t_2| \to \infty$, $\boldsymbol{P}_t^{|t_1-t_2|} \to \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\pi} \end{pmatrix}$ *where* $\boldsymbol{\pi} = (1-p, p)$. *That is, the stationary distribution of* $\boldsymbol{P}$ *is:*

$$\boldsymbol{P}^\infty = \begin{pmatrix} 1-p & p \\ 1-p & p \end{pmatrix}.$$

Proof Without loss of generality, suppose $\boldsymbol{P}$ is nontrivial, that is, $\boldsymbol{P}$ has no entries of 0 or 1. Notice that $\boldsymbol{P}$ is irreducible because $P_{uv} > 0$ for all combinations of states u and v. Next, $\boldsymbol{P}$ is aperiodic because $\Pr(x_{t_2} = u | x_{t_1} = v) > 0$ for all times $1 \le t_1 < t_2 \le n$ and all states u and v.

The stationary distribution $\boldsymbol{\pi} = (\pi_0, \pi_1)$ of $\boldsymbol{P}$ is determined by the solution to the system of equations:

$$\pi_0 = \pi_0 P_{00} + \pi_1 P_{10},$$
$$\pi_1 = \pi_0 P_{01} + \pi_1 P_{11},$$

$$1 = \pi_0 + \pi_1.$$

Substituting equations (1.21) into the equation $\pi_1 = (1-\pi_1)P_{01}+\pi_1 P_{11}$ gives one solution $\pi_1 = p$. A final substitution shows $\pi_0 = 1 - p$. ∎

The change point model discussed in this dissertation assumes the change point τ is abrupt. That is, the variables $\{x_t\}_{t=1}^{\tau}$ are independent of the variables $\{x_t\}_{t=\tau+1}^{n}$. An alternative model assumes that the change point τ maintains the one step dependence structure of the x_t variables before and after τ. The results for this model are reserved for future research.

1.4.2 The *m*-dependence Property

The m-dependence property is defined in Chung [5] and is restated as follows. A sequence of random variables $\{x_t\}_{t=1}^{n}$ is said to be *m-dependent* if $|t_1 - t_2| > m$ implies that x_{t_1} is independent of x_{t_2}. The one-step Markov dependence assumption implies that the sequence of random variables x_t is asymptotically m-dependent. Before proving this result, an interesting lemma is proved below.

Lemma 1.4.3 *Suppose that the sequence of Bernoulli(p) random variables* $\{x_t\}_{t=1}^{n}$ *follows one-step Markov dependence with transition matrix* $\boldsymbol{P}$ *as defined in equation* (1.18)*. For all* $\epsilon > 0$ *there exists an integer* m *such that for all* t_1, t_2 *satisfying* $|t_1 - t_2| > m$*,* $Cov(x_{t_1}, x_{t_2}) < \epsilon$*.*

Proof Without loss of generality, suppose $t_1 < t_2$. The covariance of x_{t_1} and x_{t_2} is:

$$\begin{aligned}
\mathrm{Cov}(x_{t_1}, x_{t_2}) &= \mathrm{E}(x_{t_1}x_{t_2}) - \mathrm{E}(x_{t_1})\mathrm{E}(x_{t_2}) \\
&= \Pr(x_{t_2} = 1 | x_{t_1} = 1)\Pr(x_{t_1} = 1) - p^2 \\
&= \boldsymbol{P}_{11}^{|t_1 - t_2|} p - p^2 \\
&= p(\boldsymbol{P}_{11}^{|t_1 - t_2|} - p).
\end{aligned}$$

Lemma 1.4.2 implies that as $|t_1 - t_2| \to \infty$, $\boldsymbol{P}_{11}^{|t_1-t_2|} \to p$. Thus, as $|t_1 - t_2| \to \infty$, $\text{Cov}(x_{t_1}, x_{t_2}) \to 0$. Therefore, for all $\epsilon > 0$ there exists an integer m such that for all t_1, t_2 satisfying $|t_1 - t_2| > m$, $\text{Cov}(x_{t_1}, x_{t_2}) < \epsilon$. ∎

Lemma 1.4.4 *Suppose that the sequence $\{x_t\}_{t=1}^n$ of Bernoulli(p) random variables follows one-step Markov dependence with transition matrix $\boldsymbol{P}$ as defined in (1.18), then $\{x_t\}_{t=1}^n$ is asymptotically m-dependent with m determined by $\boldsymbol{P}$.*

Proof Without loss of generality, suppose $1 \leq t_1 < t_2 \leq n$ and that as $n \to \infty$, t_1/n and t_2/n converge to constants in the interval $(0, 1)$, say η_1 and η_2, respectively. This forces $t_2 - t_1 \to \infty$ as $n \to \infty$. The Bernoulli distribution of both x_{t_1} and x_{t_2} leads to four possible outcomes of the joint distribution of x_{t_1} and x_{t_2}. Two of these cases are demonstrated below, as the other two are similar.

Case I: $x_{t_1} = 1$, $x_{t_2} = 1$,

$$\begin{aligned}
\text{Pr}(x_{t_1} = 1, x_{t_2} = 1) &= \text{Pr}(x_{t_2} = 1 | x_{t_1} = 1)\text{Pr}(x_{t_1} = 1) \\
&= \boldsymbol{P}_{11}^{t_2-t_1} p \\
&\to p^2 \\
&= \text{Pr}(x_{t_1} = 1)\text{Pr}(x_{t_2} = 1).
\end{aligned}$$

Case II: $x_{t_1} = 0$, $x_{t_2} = 1$,

$$\begin{aligned}
\text{Pr}(x_{t_1} = 0, x_{t_2} = 1) &= \text{Pr}(x_{t_2} = 1 | x_{t_1} = 0)\text{Pr}(x_{t_1} = 0) \\
&= \boldsymbol{P}_{01}^{t_2-t_1} (1 - p) \\
&\to p(1 - p) \\
&= \text{Pr}(x_{t_1} = 0)\text{Pr}(x_{t_2} = 1).
\end{aligned}$$

The values of the approximations of $\boldsymbol{P}_{uv}$ in both Cases I and II comes from the stationary

distribution of $\boldsymbol{P}^{\infty}$ from Lemma 1.4.2. The approximations are accurate when $t_2 - t_1 > m$, where m is dependent on ϵ as explained in Lemma 1.4.3. ∎

1.4.3 The m-dependent Central Limit Theorem

There are a variety of central limit theorems for m-dependent sequences. The first was introduced by Hoeffding and Robbins [15] and was later improved by Orey [27] for triangular arrays. The essential result used in this dissertation is given in Chung [5], Theorem 7.3.1, and it is used to show various asymptotic results for partial sums of m-dependent random variables. The result is restated below for use in later sections.

Theorem 1.4.5 *Suppose that* $\{x_n\}$ *is a sequence of m-dependent, uniformly bounded random variables such that:*

$$\frac{Var(S_n)}{n^{2/3}} \to +\infty$$

as $n \to \infty$. *Then* $\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \xrightarrow{d} N(0,1)$.

This theorem will be used several times throughout this dissertation. In Section 2.4, it will be used to show that the dependent CUSUM statistic is asymptotically normal. The consistency of the MLE will be shown in Section 3.2, along with the asymptotic distribution for the likelihood ratio statistic G_t^2 for fixed time points t.

Chapter 2

Dependent CUSUM test

The first method proposed to deal with the m-dependent sequence $y = \{x_t\}_{t=1}^n$ is a generalization of the CUSUM statistic. The dependent CUSUM (DCUSUM) statistic is similar in construction, but has added complexity in the variance and covariance due to the m-dependence assumption. A subscript of D is used throughout this chapter to denote the m-dependent calculations and differentiate the expectation, variance, and covariance from the independent case. The details of the CUSUM statistic for a fixed time t were defined in Section 1.2.1.

Similar to CUSUM, define the weighted sum S_t for a fixed time t as:

$$S_t = \sum_{j=1}^{t} x_j - \frac{t}{n}\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} a_j x_j, \quad \text{where} \quad a_j = \begin{cases} 1 - \frac{t}{n} & \text{if } 1 \le j \le t, \\ -\frac{t}{n} & \text{if } t+1 \le j \le n. \end{cases} \tag{2.1}$$

The dependent CUSUM statistic for a fixed time t is defined as:

$$\text{DCUSUM}_t = S_t/\sqrt{n}. \tag{2.2}$$

It is clear that $\text{E}(\text{DCUSUM}_t) = \text{E}_D(S_t/\sqrt{n}) = 0$. The variance calculation is given below:

$$\text{Var}(\text{DCUSUM}_t) = \text{Var}_D(S_t/\sqrt{n})$$

$$
\begin{aligned}
&= \mathrm{Var}_D\left(\sum_{j=1}^{n}\frac{a_j}{\sqrt{n}}x_j\right)\\
&= \frac{1}{n}\sum_{j=1}^{n}a_j^2\mathrm{Var}(x_j) + \frac{1}{n}\sum_{i\neq j}a_i a_j \mathrm{Cov}_D(x_i,x_j)\\
&= \frac{1}{n}\left[t\left(1-\frac{t}{n}\right)^2 + (n-t)\left(-\frac{t}{n}\right)^2\right]p(1-p) + \frac{2}{n}\sum_{i<j}a_i a_j p(\boldsymbol{P}_{11}^{|j-i|}-p)\\
&= \frac{t}{n}\left(1-\frac{t}{n}\right)p(1-p) + \frac{2}{n}\sum_{i<j}a_i a_j p(\boldsymbol{P}_{11}^{|j-i|}-p)\\
&= p\left[\frac{t}{n}\left(1-\frac{t}{n}\right)(1-p) + \frac{2}{n}\sum_{i<j}a_i a_j(\boldsymbol{P}_{11}^{|j-i|}-p)\right]\\
&= \sigma^2_{D,t}.
\end{aligned}
$$

Under the assumption of m-dependence, assuming m is known and $i<j$, the variance can be reduced to:

$$
\mathrm{Var}(\mathrm{DCUSUM}_t) = p\left[\frac{t}{n}\left(1-\frac{t}{n}\right)(1-p) + \frac{2}{n}\sum_{0<j-i\leq m}a_i a_j(\boldsymbol{P}_{11}^{j-i}-p)\right]. \tag{2.3}
$$

2.1 Variance of DCUSUM$_t$ Under m-dependence

The variance of DCUSUM$_t$ obviously depends on the values of t and m. In the independent case, the asymptotic distribution of the maximal CUSUM statistic is determined over a range of values $l \leq t/n \leq h$ where $l, h \in (0,1)$. The choice of l and h is discussed by Miller and Siegmund [26], and aims to improve performance of the testing methods by removing time points from the boundaries of the sequence that lead to inflated test statistics. A similar technique is applied to determine the asymptotic distribution of the maximal DCUSUM statistic.

Assuming that $i<j$, we can split the sum:

$$
\sum_{0<j-i\leq m}a_i a_j(\boldsymbol{P}_{11}^{j-i}-p) = \sum_{0<j-i\leq m}c_{ij},
$$

from (2.3) into three parts as follows:

$$\begin{aligned}\sum_{0<j-i\leq m} c_{ij} &= \sum_{\substack{1\leq i<j\leq t \\ 0<j-i\leq m}} c_{ij} + \sum_{\substack{i\leq t<j \\ 0<j-i\leq m}} c_{ij} + \sum_{\substack{t+1\leq i<j\leq n \\ 0<j-i\leq m}} c_{ij} \\ &= \quad A \quad + \quad B \quad + \quad C. \end{aligned} \tag{2.4}$$

This leads to four possible cases for the value of m:

$$\begin{aligned} \text{Case I}: &\quad m \leq \min(t, n-t), \\ \text{Case II}: &\quad t < m \leq n-t, \\ \text{Case III}: &\quad n-t < m \leq t, \\ \text{Case IV}: &\quad m > \max(t, n-t). \end{aligned}$$

Recall that the m in m-dependence is independent of the time points t and n. It is assumed that as $n \to \infty$ we have $t/n \to \eta$ where $\eta \in (l, h)$. This implies that as $n \to \infty$, $t \to \infty$, which forces the value of m to satisfy $m \leq \min(t, n-t)$. Therefore, all cases reduce to Case I. Hence, only Case I is necessary to discuss when exploring the asymptotic distribution of the DCUSUM statistic. The representation of the sum (2.4) for this case is discussed below.

Consider the matrix of pairs of time points with rows representing the index i and columns the index j for the coefficients c_{ij} in the sum (2.4). The vertical and horizontal lines represent when the coefficients a_i change from $(1 - t/n)$ to $-t/n$.

$$
\begin{array}{c|c}
i \Big/ j & \begin{array}{cccc|ccc|c} 1 & 2 & \cdots & t & t{+}1 & \cdots & n{-}1 & n \end{array} \\
\hline
\begin{array}{c} 1 \\ 2 \\ \vdots \\ t \\ \hline t{+}1 \\ \vdots \\ n{-}1 \\ n \end{array} &
\left(\begin{array}{cccc|ccc|c}
c_{11} & c_{12} & \cdots & c_{1t} & c_{1\,t+1} & \cdots & c_{1\,n-1} & c_{1n} \\
c_{21} & c_{22} & \cdots & c_{2t} & c_{2\,t+1} & \cdots & c_{2\,n-1} & c_{2n} \\
\vdots & \vdots & \cdots & \vdots & \vdots & \cdots & & \vdots \\
c_{t1} & c_{t2} & \cdots & c_{tt} & c_{t\,t+1} & \cdots & c_{t\,n-1} & c_{tn} \\
\hline
c_{t+1\,1} & c_{t+1\,2} & \cdots & c_{t+1\,t} & c_{t+1\,t+1} & \cdots & c_{t+1\,n-1} & c_{t+1\,n} \\
\vdots & \vdots & \cdots & \vdots & \vdots & \cdots & & \vdots \\
c_{n-1\,1} & c_{n-1\,2} & \cdots & c_{n-1\,t} & c_{n-1\,t+1} & \cdots & c_{n-1\,n-1} & c_{n-1\,n} \\
c_{n1} & c_{n2} & \cdots & c_{nt} & c_{n\,t+1} & \cdots & c_{n\,n-1} & c_{nn}
\end{array}\right)
\end{array}
\tag{2.5}
$$

The four sections of this matrix can be identified as $\left(\begin{array}{c|c} A & B \\ \hline 0 & C \end{array}\right)$. Notice that the lower left block of the matrix is zero because of the assumption that $i < j$. Coefficients are counted starting with the diagonal entries with subscripts $c_{i\,i+1}$. For A, there are $t-1$ of these coefficients, for B there is 1, and for C there are $n-t-1$. Next, the number of $c_{i\,i+2}$ entries are counted. There are $t-2$ of these in A, 2 in B, and $n-t-2$ in C. This process continues until the difference in subscripts $j-i > m$. The total counts of coefficients for each of A, B, and C for a fixed value of m are given below:

$$
\begin{aligned}
A &= \left(1-\frac{t}{n}\right)^2 \sum_{w=1}^{m} (t-w)\left(\boldsymbol{P}_{11}^{w} - p\right), \\
B &= \left(-\frac{t}{n}\right)\left(1-\frac{t}{n}\right) \sum_{w=1}^{m} w\left(\boldsymbol{P}_{11}^{w} - p\right), \\
C &= \left(-\frac{t}{n}\right)^2 \sum_{w=1}^{m} (n-t-w)\left(\boldsymbol{P}_{11}^{w} - p\right).
\end{aligned}
\tag{2.6}
$$

2.1.1 Asymptotic Value of Var(DCUSUM$_t$)

The limiting value of Var(DCUSUM$_t$) is discussed below. Recall the assumption that $t/n \to \eta \in (0,1)$ as $n \to \infty$.

$$\begin{aligned}
\text{Var}(\text{DCUSUM}_t) &= \sigma^2_{D,t} \\
&= p\left[\frac{t}{n}\left(1-\frac{t}{n}\right)(1-p) + 2\sum_{0<j-i\le m}\frac{c_{ij}}{n}\right] \\
&= p\left[\frac{t}{n}\left(1-\frac{t}{n}\right)(1-p) + 2\left(\left(1-\frac{t}{n}\right)^2\sum_{w=1}^{m}\frac{t-w}{n}(\boldsymbol{P}_{11}^w - p)\right.\right. \\
&\qquad + \left(-\frac{t}{n}\right)\left(1-\frac{t}{n}\right)\sum_{w=1}^{m}\frac{w}{n}(\boldsymbol{P}_{11}^w - p) \\
&\qquad \left.\left. + \left(-\frac{t}{n}\right)^2\sum_{w=1}^{m}\frac{n-t-w}{n}(\boldsymbol{P}_{11}^w - p)\right)\right] \\
&\to p\left[\eta(1-\eta)(1-p) + 2\left((1-\eta)^2\eta\sum_{i=1}^{m}(\boldsymbol{P}_{11}^w - p)\right.\right. \\
&\qquad \left.\left. + (1-\eta)(-\eta)^2\sum_{w=1}^{m}(\boldsymbol{P}_{11}^w - p)\right)\right] \\
&= p\eta(1-\eta)\left((1-p) + 2\sum_{w=1}^{m}(\boldsymbol{P}_{11}^w - p)\right).
\end{aligned}$$

2.2 Covariance of DCUSUM$_{t_1}$ and DCUSUM$_{t_2}$

Define the statistic $T_t = \text{DCUSUM}_t/\hat{\sigma}_{D,t}$, where $\hat{\sigma}_{D,t}$ is any consistent estimator of $\sigma_{D,t}$. For the purposes of this dissertation, define:

$$\begin{aligned}
\hat{\sigma}_{D,t} = \hat{p}\left[\frac{t}{n}\left(1-\frac{t}{n}\right)(1-\hat{p}) + 2\left(\left(1-\frac{t}{n}\right)^2\sum_{w=1}^{m}\frac{t-w}{n}(\hat{\boldsymbol{P}}_{11}^w - \hat{p})\right.\right. \\
+ \left(-\frac{t}{n}\right)\left(1-\frac{t}{n}\right)\sum_{w=1}^{m}\frac{w}{n}(\hat{\boldsymbol{P}}_{11}^w - \hat{p}) \\
\left.\left. + \left(-\frac{t}{n}\right)^2\sum_{w=1}^{m}\frac{n-t-w}{n}(\hat{\boldsymbol{P}}_{11}^w - \hat{p})\right)\right],
\end{aligned}$$

where $\hat{p}$ and $\hat{\boldsymbol{P}}_{11}$ are defined by (1.2) and (3.7), respectively. The estimate $\hat{p}$ of p is consistent by Theorem 1.4.5, and $\hat{\boldsymbol{P}}_{11}$ is consistent by Corollary 3.2.6. The test statistic for detecting a change point τ is:

$$T^2_{\max} = \max_t T^2_t. \tag{2.7}$$

The covariance between DCUSUM_{t_1} and DCUSUM_{t_2} for any two times $t_1 < t_2$ is necessary to obtain the asymptotic distribution of the $T^2_{\max}$ statistic, as well as an approximate Worsley type upper bound for tail probabilities. Let a_i be the coefficients for DCUSUM_{t_1} and b_j be the coefficients for DCUSUM_{t_2}, then:

$$\begin{aligned}
\text{Cov}(\text{DCUSUM}_{t_1}, \text{DCUSUM}_{t_2}) &= \text{Cov}(S_{t_1}/\sqrt{n}, S_{t_2}/\sqrt{n}) \\
&= \text{Cov}\left(\left[\sum_{i=1}^{n} \frac{a_i}{\sqrt{n}} X_i\right]\left[\sum_{j=1}^{n} \frac{b_j}{\sqrt{n}} X_j\right]\right) \\
&= \text{Cov}\left(\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\left(\frac{a_j}{\sqrt{n}} X_i \frac{b_j}{\sqrt{n}} X_j\right)\right]\right) \\
&= \sum_{i=1}^{n}\left[\sum_{j=1}^{n} \frac{a_i b_j}{n} \text{Cov}(X_i X_j)\right] \\
&= \sum_{i=1}^{n} \frac{a_i b_j}{n} \text{Var}(X_i) + \sum_{i \neq j} \frac{a_i b_j}{n} \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \frac{a_i b_j}{n} p(1-p) + \sum_{i \neq j} \frac{a_i b_j}{n} p(\boldsymbol{P}_{11}^{|j-i|} - p).
\end{aligned}$$

Under the assumption of m-dependence, a similar simplification in the covariance occurs:

$$\text{Cov}(\text{DCUSUM}_{t_1}, \text{DCUSUM}_{t_2}) = \frac{p}{n}\left[\sum_{i=1}^{n} a_i b_i (1-p) + \sum_{0<|j-i|\leq m} a_i b_j (\boldsymbol{P}_{11}^{|j-i|} - p)\right]. \tag{2.8}$$

Note that the m in m-dependence is independent of the time points t_1, t_2, and n. Recall the assumption that as $n \to \infty$ both $t_1/n \to \eta_1$ and $t_2/n \to \eta_2$ where $\eta_1, \eta_2 \in (l, h)$. This implies that as $n \to \infty$, all of $t_1 \to \infty$, $t_2 \to \infty$, and $t_2 - t_1 \to \infty$. Even though the limit of

$t_2 - t_1$ grows to infinity, the cases where $t_2 - t_1 < m$ are included for implementation into the DCUSUM test algorithm described in Chapter 4. The possible values of m are limited to the following categories:

$$\begin{aligned}
1: &\quad m \le t_2 - t_1 \le t_1 \le n - t_2, \\
2: &\quad m \le t_1 \le t_2 - t_1 \le n - t_2, \\
3: &\quad m \le t_1 \le n - t_2 \le t_2 - t_1, \\
4: &\quad m \le t_2 - t_1 \le n - t_2 \le t_1, \\
5: &\quad m \le n - t_2 \le t_2 - t_1 \le t_1, \\
6: &\quad m \le n - t_2 \le t_1 \le t_2 - t_1, \\
7: &\quad t_2 - t_1 \le m \le t_1 \le n - t_2, \\
8: &\quad t_2 - t_1 \le m \le n - t_2 \le t_1.
\end{aligned}$$

The resulting covariance, $\mathrm{Cov}(\mathrm{DCUSUM}_{t_1}, \mathrm{DCUSUM}_{t_2})$, of these eight cases reduces to two unique possibilities:

$$\begin{aligned}
&\text{Case I (1 - 6)}: \quad m \le \min(t_1, t_2 - t_1, n - t_2), \\
&\text{Case II (7, 8)}: \quad t_2 - t_1 \le m \le \min(t_1, n - t_2).
\end{aligned}$$

For both cases, the sum:

$$\sum_{0<|j-i|\le m} a_i b_j p(\boldsymbol{P}_{11}^{|j-i|} - p) = \sum_{0<|j-i|\le m} c_{ij},$$

from equation (2.8) may be split into four parts. Define $a = (1 - t_1/n), a' = t_1/n, b = (1 - t_2/n)$ and $b' = t_2/n$. Then:

$$\sum_{0<|j-i|\le m} c_{ij} = ab \sum_{i=1}^{m} c_{i,ab} + a'b \sum_{i=1}^{m} c_{i,a'b} + ab' \sum_{i=1}^{m} c_{i,ab'} + a'b' \sum_{i=1}^{m} c_{i,a'b'}$$

$$= \quad AB \quad + \quad A'B \quad + \quad AB' \quad + \quad A'B'. \tag{2.9}$$

Similar to the matrix (2.5), consider the matrix of pairs of time points with rows representing the index i of a_i corresponding to t_1 and columns the index j of b_j corresponding to t_2 for the coefficients c_{ij} in the sum (2.9). The vertical and horizontal lines represent when the coefficients a_i or b_j change from $(1 - t_1/n)$ or $(1 - t_2/n)$ to $-t_1/n$ or $-t_2/n$ respectively.

a_i/b_j	1	2	$\cdots$	t_1	$t_1{+}1$	$\cdots$	t_2	$t_2{+}1$	$\cdots$	$n{-}1$	n
1	c_{11}	c_{12}	$\cdots$	c_{1t_1}	$c_{1\,t_1+1}$	$\cdots$	c_{1t_2}	$c_{1\,t_2+1}$	$\cdots$	$c_{1\,n-1}$	c_{1n}
2	c_{21}	c_{22}	$\cdots$	c_{2t_1}	$c_{2\,t_1+1}$	$\cdots$	c_{2t_2}	$c_{2\,t_2+1}$	$\cdots$	$c_{2\,n-1}$	c_{2n}
$\vdots$	$\vdots$		$\cdots$	$\vdots$		$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	
t_1	$c_{t_1 1}$	$c_{t_1 2}$	$\cdots$	$c_{t_1 t_1}$	$c_{t_1\,t_1+1}$	$\cdots$	$c_{t_1 t_2}$	$c_{t_1\,t_2+1}$	$\cdots$	$c_{t_1\,n-1}$	$c_{t_1 n}$
$t_1{+}1$	$c_{t_1+1\,1}$	$c_{t_1+1\,2}$	$\cdots$	$c_{t_1+1\,t_1}$	$c_{t_1+1\,t_1+1}$	$\cdots$	$c_{t_1+1\,t_2}$	$c_{t_1+1\,t_2+1}$	$\cdots$	$c_{t_1+1\,n-1}$	$c_{t_1+1\,n}$
$\vdots$	$\vdots$		$\cdots$	$\vdots$		$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	
t_2	$c_{t_2 1}$	$c_{t_2 2}$	$\cdots$	$c_{t_2 t_1}$	$c_{t_2\,t_1+1}$	$\cdots$	$c_{t_2 t_2}$	$c_{t_2\,t_2+1}$	$\cdots$	$c_{t_2\,n-1}$	$c_{t_2 n}$
$t_2{+}1$	$c_{t_2+1\,1}$	$c_{t_2+1\,2}$	$\cdots$	$c_{t_2+1\,t_1}$	$c_{t_2+1\,t_1+1}$	$\cdots$	$c_{t_2+1\,t_2}$	$c_{t_2+1\,t_2+1}$	$\cdots$	$c_{t_2+1\,n-1}$	$c_{t_2+1\,n}$
$\vdots$	$\vdots$		$\cdots$	$\vdots$		$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	
$n{-}1$	$c_{n-1\,1}$	$c_{n-1\,2}$	$\cdots$	$c_{n-1\,t_1}$	$c_{n-1\,t_1+1}$	$\cdots$	$c_{n-1\,t_2}$	$c_{n-1\,t_2+1}$	$\cdots$	$c_{n-1\,n-1}$	$c_{n-1\,n}$
n	c_{n1}	c_{n2}	$\cdots$	c_{nt_1}	$c_{n\,t_1+1}$	$\cdots$	c_{nt_2}	$c_{n\,t_2+1}$	$\cdots$	$c_{n\,n-1}$	c_{nn}

The four sections of this matrix can be identified as $\left(\begin{array}{c|c} AB & AB' \\ \hline A'B & A'B' \end{array}\right)$. The counting of coefficients using this matrix will change depending on the case considered. Both Cases I and II are discussed below.

2.2.1 Coefficients for Case I: $m \leq \min(t_1, t_2 - t_1, n - t_2)$

Under the assumptions of Case I, the upper right block of the matrix AB' is zero because of the assumption that $m < t_2 - t_1$ and the fact that all entries in that block are separated by at least $t_2 - t_1$ time points.

Coefficients in the other three blocks are counted above and below the diagonal starting with with subscripts $c_{i\,i+1}$, then $c_{i-1\,i}$. For AB, there are t_1 of these coefficients above the diagonal and $t_1 - 1$ below the diagonal for a total of $2t_1 - 1$ coefficients. For $A'B$ there are $t_2 - t_1 - 1$ of these coefficients above the diagonal and $t_2 - t_1 + 1$ below the diagonal for a total of $2(t_2 - t_1)$ coefficients. For $A'B'$ there are $n - t_2$ of these coefficients above the diagonal and $n - t_2 - 1$ below the diagonal for a total of $2(n - t_2) - 1$ coefficients.

Next, the number of $c_{i\,i+2}$ and $c_{i-2\,i}$ coefficients are counted. In AB, there are t_1 of these coefficients above the diagonal and $t_1 - 2$ below the diagonal for a total of $2t_1 - 2$ coefficients. In $A'B$, there are $t_2 - t_1 - 2$ of these coefficients above the diagonal and $t_2 - t_1 + 2$ below the diagonal for a total of $t_2 - t_1$ coefficients. In $A'B'$, there are $n - t_2$ of these coefficients above the diagonal and $n - t_2 - 2$ below the diagonal for a total of $2(n - t_2) - 2$ coefficients.

This process continues until the difference in subscripts $j - i > m$. The total counts of coefficients for each of $AB, A'B$, and $A'B'$ from (2.9) for a fixed value of m are given below:

$$\begin{aligned}
AB &= \left(1 - \frac{t_1}{n}\right)\left(1 - \frac{t_2}{n}\right)\sum_{w=1}^{m}(2t_1 - w)\left(\boldsymbol{P}_{11}^{w} - p\right), \\
A'B &= \left(-\frac{t_1}{n}\right)\left(1 - \frac{t_2}{n}\right)\sum_{w=1}^{m}[2(t_2 - t_1)]\left(\boldsymbol{P}_{11}^{w} - p\right), \\
AB' &= 0, \\
A'B' &= \left(-\frac{t_1}{n}\right)\left(-\frac{t_2}{n}\right)\sum_{w=1}^{m}[2(n - t_2) - w]\left(\boldsymbol{P}_{11}^{w} - p\right). \qquad (2.10)
\end{aligned}$$

2.2.2 Coefficients for Case II: $t_2 - t_1 \leq m \leq \min(t_1, n - t_2)$

Under the assumptions of Case II, all four blocks of the coefficient matrix must be counted. Notice that the number of coefficients in $AB, A'B$, and $A'B'$ are similar to those from Case I, with an extra term. This extra term is due to the fact that when the value of m is larger than $t_2 - t_1$, more terms have a nonzero covariance. The details of the counting are omitted, but the strategy is the same as in Case I. The counts for Case II from the sum (2.9) are

given below:

$$
\begin{aligned}
AB &= \left(1-\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)\left[\sum_{w=1}^{t_2-t_1}(2t_1-w)\left(\boldsymbol{P}_{11}^{w}-p\right)\right. \\
&\qquad\left. +\sum_{w=t_2-t_1+1}^{m}(t_2+t_1-2w)\left(\boldsymbol{P}_{11}^{w}-p\right)\right], \\
A'B &= \left(-\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)\left[\sum_{w=1}^{t_2-t_1}[2(t_2-t_1)]\left(\boldsymbol{P}_{11}^{w}-p\right)\right. \\
&\qquad\left. +\sum_{w=t_2-t_1+1}^{m}(t_2-t_1+w)\left(\boldsymbol{P}_{11}^{w}-p\right)\right], \\
AB' &= \left(1-\frac{t_1}{n}\right)\left(-\frac{t_2}{n}\right)\sum_{w=t_2-t_1+1}^{m}[w-(t_2-t_1)]\left(\boldsymbol{P}_{11}^{w}-p\right), \\
A'B' &= \left(-\frac{t_1}{n}\right)\left(-\frac{t_2}{n}\right)\left[\sum_{w=1}^{t_2-t_1}[2(n-t_2)-w]\left(\boldsymbol{P}_{11}^{w}-p\right)\right. \\
&\qquad\left. +\sum_{w=t_2-t_1+1}^{m}[2(n-w)-(t_2+t_1)]\left(\boldsymbol{P}_{11}^{w}-p\right)\right]. \qquad (2.11)
\end{aligned}
$$

2.2.3 Asymptotic value of $\mathbf{Cov}(\mathbf{DCUSUM}_{t_1}, \mathbf{DCUSUM}_{t_2})$

The limiting value of $\mathrm{Cov}(\mathrm{DCUSUM}_{t_1}, \mathrm{DCUSUM}_{t_2})$ for Case I is discussed below. This is sufficient because of the fact that $t_2 - t_1 \to \infty$. Because m is fixed, as $n \to \infty$, $m < t_2 - t_1$. Therefore, Case II is not possible as $n \to \infty$. Recall the assumption that $t_1/n \to \eta_1$ and $t_2/n \to \eta_2$ where $0 < \eta_1 < \eta_2 < 1$.

$$\text{Case I}: \quad m \le \min(t_1, t_2 - t_1, n - t_2).$$

$$
\begin{aligned}
\mathrm{Cov}(\mathrm{DCUSUM}_{t_1}, \mathrm{DCUSUM}_{t_2}) &= p\left[\sum_{i=1}^{n}\frac{a_i b_i}{n}(1-p) + \sum_{0\le|j-i|\le m}\frac{a_i b_j}{n}(\boldsymbol{P}_{11}^{|j-i|}-p)\right] \\
&\to p\left[(1-p)\eta_1(1-\eta_2) + 2\eta_1(1-\eta_2)\sum_{w=1}^{m}(\boldsymbol{P}_{11}^{w}-p)\right] \\
&= \eta_1(1-\eta_2)p\left[1-p+2\sum_{w=1}^{m}(\boldsymbol{P}_{11}^{w}-p)\right].
\end{aligned}
$$

Combining the results in this section with those in Section 2.1 leads to the following asymptotic covariance of T_{t_1} and T_{t_2}:

$$\begin{aligned}\mathrm{Cov}(T_{t_1}, T_{t_2}) &= \frac{\mathrm{Cov}(\mathrm{DCUSUM}_{t_1}, \mathrm{DCUSUM}_{t_2})}{\hat{\sigma}_{D,t_1}\hat{\sigma}_{D,t_2}} \\ &\to \frac{\eta_1(1-\eta_2)p\left[1-p+2\sum_{w=1}^{m}(\boldsymbol{P}_{11}^{w}-p)\right]}{\sqrt{p\eta_1(1-\eta_1)\left((1-p)+2\sum_{w=1}^{m}(\boldsymbol{P}_{11}^{i}-p)\right)}\sqrt{p\eta_2(1-\eta_2)\left((1-p)+2\sum_{w=1}^{m}(\boldsymbol{P}_{11}^{w}-p)\right)}} \\ &= \left(\frac{\eta_1(1-\eta_2)}{(1-\eta_1)\eta_2}\right)^{1/2}. \end{aligned} \tag{2.12}$$

2.3 Asymptotic Distribution of the Maximum DCUSUM Statistic

Recall from Section 1.2.1 that the independent statistic $T^2_{\max}$ is asymptotically the sum of squared Brownian Bridge processes. The asymptotic distribution for the $T^2_{\max}$ statistic for the DCUSUM case is summarized in the theorem below. The calculations of the limiting values of the mean, variance, and covariance of T_t are provided in the previous subsections.

Theorem 2.3.1 *Suppose $\{x_t\}_{t=1}^{n}$ is an m-dependent sequence of Bernoulli(p) random variables with one step Markov dependence defined by the transition matrix $\boldsymbol{P}$, and the value of m is known. The test statistic:*

$$T^2_{\max} = \max_{l\le t/n\le h} T_t^2 \xrightarrow{D} \sup_{l\le\eta\le h}\frac{B^2(\eta)}{\eta(1-\eta)},$$

where $B(\eta)$ is a Brownian bridge process on the interval $[0,1]$.

Proof Define S_t as in equation (2.1), then the random variables a_jx_j that make up S_t are uniformly bounded by 1. From equation (2.3), the a consistent estimate of the variance of S_t is given as:

$$\widehat{\mathrm{Var}}_D(S_t) = \hat{p}\left[t\left(1-\frac{t}{n}\right)(1-\hat{p}) + 2\sum_{0<j-i<m} a_ia_j(\hat{\boldsymbol{P}}_{11}^{|j-i|}-\hat{p})\right].$$

As n tends to infinity:

$$\begin{aligned}
\widehat{\mathrm{Var}}_D(S_t)/n^{2/3} &= \hat{p}\left[\frac{t}{n^{2/3}}\left(1-\frac{t}{n}\right)(1-\hat{p}) + \frac{2}{n^{2/3}}\sum_{0<j-i<m} a_i a_j(\hat{\boldsymbol{P}}_{11}^{|j-i|} - \hat{p})\right] \\
&= n^{1/3}\hat{p}\left[\frac{t}{n}\left(1-\frac{t}{n}\right)(1-\hat{p}) + \frac{2}{n}\sum_{0<j-i<m} a_i a_j(\hat{\boldsymbol{P}}_{11}^{|j-i|} - \hat{p})\right] \\
&\to \infty.
\end{aligned}$$

By Theorem 1.4.5 and the consistency of $\hat{\sigma}_{D,t}$:

$$T_t = \frac{\mathrm{DCUSUM}_t}{\hat{\sigma}_{D,t}} = \frac{S_t}{\sqrt{\widehat{\mathrm{Var}}_D(S_t)}} = \frac{S_t - \mathrm{E}_D(S_t)}{\sqrt{\widehat{\mathrm{Var}}_D(S_t)}} \xrightarrow{D} \mathrm{N}(0,1).$$

For the independent T_t statistic, an application of the traditional CLT gives the convergence $T_t \xrightarrow{D} \mathrm{N}(0,1)$.

Both of the test statistics T_t resulting from the independent and dependent assumptions have asymptotically normal distributions. The normal distribution is completely determined by the mean, variance, and covariance. If the means, variances, and covariances in both cases are asymptotically equivalent, then the asymptotic results for the statistic $T^2_{\max}$ in either case will be equivalent.

A comparison of the asymptotic results of Section 1.2.1 and Chapter 2 is given below. The subscript D denotes the calculations for the DCUSUM statistic.

$$\begin{aligned}
\mathrm{E}(T_t) &= 0 = \mathrm{E}_D(T_t), \\
\mathrm{Var}(T_t) &= 1 = \mathrm{Var}_D(T_t), \\
\mathrm{Cov}(T_{t_1}, T_{t_2}) &\approx \left(\frac{\eta_1(1-\eta_2)}{(1-\eta_1)\eta_2}\right)^{1/2} \approx \mathrm{Cov}_D(T_{t_1}, T_{t_2}).
\end{aligned}$$

Notice that the means, variances, and asymptotic covariances (2.12) and (1.6) are identical. Therefore, the independent and dependent T_t^2 statistics will have the exact same

asymptotic behavior. Hence, the asymptotic distribution of the statistic $T^2_{\max}$ is the same for both the independent CUSUM and m-dependent DCUSUM test. ∎

Approximate p-values for the DCUSUM test can be found by applying Theorem 2.3.1 to $T^2_{\max}$ and using (1.14), which is restated below for convenience:

$$\Pr\left(\sup_{l\le\eta\le h}\frac{B^2(\eta)}{\eta(1-\eta)}\ge T\right)\approx \left(\frac{Te^{-T}}{2\pi}\right)^{1/2} \times\left[\left(1-\frac{1}{T}\right)\log\left(\frac{(1-l)h}{l(1-h)}\right)+\frac{4}{T}+O\left(\frac{1}{T^2}\right)\right].$$

The asymptotic results rely on the fact that the value m is known. In practice, this value is unknown and must be estimated from the data. This procedure is discussed in Section 4.1.

2.4 Upper Bound for DCUSUM Tail Probabilities

An alternative approach to the p-value approximation of the DCUSUM statistic using (1.14) is to use a Worsley type upper bound for the p-value as mentioned in Section 1.3.2. In particular, the upper bound (1.16) is applied to the statistic $T^2_{\max}$ and is calculated as follows:

$$\begin{aligned}\Pr(T^2_{\max}>T)&\le\sum_{t=nl}^{nh}\Pr(T_t^2>T)-\sum_{t=nl}^{nh-1}\Pr\left(\{T_t^2>T\}\cap\{T_{t+1}^2>T\}\right)\\&=\sum_{t=nl}^{nh}\Pr(T_t^2>T)-\Bigg[\sum_{t=nl}^{nh-1}\Pr\left(\{T_t>\sqrt{T}\}\cap\{T_{t+1}>\sqrt{T}\}\right)\\&\qquad+\Pr\left(\{T_t>\sqrt{T}\}\cap\{T_{t+1}<-\sqrt{T}\}\right)\\&\qquad+\Pr\left(\{T_t<-\sqrt{T}\}\cap\{T_{t+1}>\sqrt{T}\}\right)\\&\qquad+\Pr\left(\{T_t<-\sqrt{T}\}\cap\{T_{t+1}<-\sqrt{T}\}\right)\Bigg].\end{aligned}\tag{2.13}$$

In order to calculate the probabilities in equation (2.13), the distribution of T_t^2 and

the joint distribution of T_t and T_{t+1} must be calculated. The following lemmas give the asymptotic distribution of T_t^2 as χ_1^2 and asymptotic joint distribution of T_t and T_{t+1} as bivariate normal.

Lemma 2.4.1 *Suppose the sequence $\{x_t\}_{t=1}^n$ is m-dependent, and the value of m is known. The statistic:*

$$T_t^2 = \left(\frac{DCUSUM_t}{\hat{\sigma}_{D,t}}\right)^2,$$

has asymptotic distribution χ_1^2.

Proof From the proof of Theorem 2.3.1:

$$T_t \xrightarrow{D} \mathrm{N}(0,1).$$

Therefore:

$$T_t^2 = \left(\frac{\mathrm{DCUSUM}_t}{\hat{\sigma}_{D,t}}\right)^2 \xrightarrow{D} \chi_1^2.$$

∎

In order to show the asymptotic joint distribution of T_{t_1} and T_{t_2} is bivariate normal, it must be shown that any linear combination $S_{t_1,t_2} = cS_{t_1} + dS_{t_2},$ for any $c, d \in \mathbb{R}$ is normally distributed. Define the partial sum:

$$S_{t_1,t_2} = \sum_{j=1}^{n} \gamma_j x_j, \quad \text{where} \quad \gamma_j = \begin{cases} c\left(1-\frac{t_1}{n}\right) + d\left(1-\frac{t_2}{n}\right) & \text{if } 1 \le j \le t_1, \\ c\left(-\frac{t_1}{n}\right) + d\left(1-\frac{t_2}{n}\right) & \text{if } t_1+1 \le j \le t_2, \\ c\left(-\frac{t_1}{n}\right) + d\left(-\frac{t_2}{n}\right) & \text{if } t_2+1 \le j \le n. \end{cases} \tag{2.14}$$

The variance of S_{t_1,t_2} is given as:

$$\begin{aligned} \mathrm{Var}(S_{t_1,t_2}) &= \sum_{i=1}^{n} \gamma_i^2 \mathrm{Var}(x_j) + 2 \sum_{0<j-i<m} \gamma_i \gamma_j \mathrm{Cov}(x_i, x_j) \\ &= \sum_{i=1}^{n} \gamma_i^2 p(1-p) + 2 \sum_{0<j-i<m} \gamma_i \gamma_j \left(\boldsymbol{P}_{11}^{|j-i|} - p\right). \end{aligned} \tag{2.15}$$

Table 2.1: Coefficient Locations and Counts

Location	Count
$\{1 \leq i < j \leq t_1\}$	$\gamma_1^2 \sum_{w=1}^{m} (t_1 - w) (\boldsymbol{P}_{11}^w - p)$
$\{i \leq t_1 < j \leq t_2\}$	$\gamma_1 \gamma_2 \sum_{w=1}^{m} w (\boldsymbol{P}_{11}^w - p)$
$\{t_1 + 1 \leq i < j \leq t_2\}$	$\gamma_2^2 \sum_{w=1}^{m} (t_2 - t_1 - w) (\boldsymbol{P}_{11}^w - p)$
$\{t_1 + 1 \leq i \leq t_2 < j \leq n\}$	$\gamma_2 \gamma_3 \sum_{w=1}^{m} w (\boldsymbol{P}_{11}^w - p)$
$\{t_2 + 1 \leq i < j \leq n\}$	$\gamma_3^2 \sum_{w=1}^{m} (n - t_2 - w) (\boldsymbol{P}_{11}^w - p)$

The coefficients of the second term can be counted using a similar counting technique for the coefficients (2.6), except with five categories. The coefficient counts for each category are given in Table 2.1. For notational purposes, define:

$$\begin{aligned}
\gamma_1 &= c\left(1 - \frac{t_1}{n}\right) + d\left(1 - \frac{t_2}{n}\right), \\
\gamma_2 &= c\left(1 - \frac{t_1}{n}\right) + d\left(-\frac{t_2}{n}\right), \\
\gamma_3 &= c\left(-\frac{t_1}{n}\right) + d\left(-\frac{t_2}{n}\right).
\end{aligned}$$

The asymptotic variance of $S_{t_1,t_2}/\sqrt{n}$ is calculated below, separated into three steps. The first term in the variance (2.15) is calculated first, then the second term is calculated, and finally, the two are combined and the limit is taken.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \gamma_i^2 \mathrm{Var}(x_j) &= \frac{1}{n}\sum_{i=1}^{n} \gamma_i^2 p(1-p) \\
&= \frac{p(1-p)}{n}\left\{ t_1 \left[c\left(1 - \frac{t_1}{n}\right) + d\left(1 - \frac{t_2}{n}\right)\right]^2 \right. \\
&\qquad + (t_2 - t_1)\left[c\left(-\frac{t_1}{n}\right) + d\left(1 - \frac{t_2}{n}\right)\right]^2 \\
&\qquad \left. + (n - t_2)\left[c\left(-\frac{t_1}{n}\right) + d\left(-\frac{t_2}{n}\right)\right]^2 \right\}
\end{aligned}$$

$$= \frac{p(1-p)}{n}\left\{c^2 t_1\left(1-\frac{t_1}{n}\right) + 2cdt_1\left(1-\frac{t_2}{n}\right) + d^2 t_2\left(1-\frac{t_2}{n}\right)\right\}$$
$$\to p(1-p)\left\{c^2\eta_1(1-\eta_1) + 2cd\eta_1(1-\eta_2) + d^2\eta_2(1-\eta_2)\right\}. \tag{2.16}$$

$$\begin{aligned}
\frac{2}{n}\sum_{0<j-i<m}\gamma_i\gamma_j \mathrm{Cov}(x_i,x_j) &= \frac{2}{n}p\sum_{0<j-i<m}\gamma_i\gamma_j\left(\boldsymbol{P}_{11}^{|j-i|}-p\right)\\
&= \frac{2}{n}p\left\{\gamma_1^2\sum_{w=1}^{m}(t_1-w)\left(\boldsymbol{P}_{11}^{w}-p\right) + \gamma_1\gamma_2\sum_{w=1}^{m}w\left(\boldsymbol{P}_{11}^{w}-p\right)\right.\\
&\quad + \gamma_2^2\sum_{w=1}^{m}(t_2-t_1-w)\left(\boldsymbol{P}_{11}^{w}-p\right) + \gamma_2\gamma_3\sum_{w=1}^{m}w\left(\boldsymbol{P}_{11}^{w}-p\right)\\
&\quad \left. + \gamma_3^2\sum_{w=1}^{m}(n-t_2-w)\left(\boldsymbol{P}_{11}^{w}-p\right)\right\}\\
&\to 2p\left\{\left[c(1-\eta_1)+d(1-\eta_2)\right]^2\eta_1\right.\\
&\quad + \left[c(1-\eta_1)+d(-\eta_2)\right]^2(\eta_2-\eta_1)\\
&\quad \left. + \left[c(-\eta_1)+d(-\eta_2)\right]^2(1-\eta_2)\right\}\sum_{w=1}^{m}\left(\boldsymbol{P}_{11}^{w}-p\right)\\
&= 2p\left\{c^2\eta_1(1-\eta_1) + 2cd\eta_1(1-\eta_2)\right.\\
&\quad \left. + d^2\eta_2(1-\eta_2)\right\}\sum_{w=1}^{m}\left(\boldsymbol{P}_{11}^{w}-p\right).
\end{aligned} \tag{2.17}$$

Notice that both terms (2.16) and (2.17) in the sum of the asymptotic variance of S_{t_1,t_2} have a common constant. Combining these yields the total asymptotic variance:

$$\begin{aligned}
\mathrm{Var}\left(\frac{S_{t_1,t_2}}{\sqrt{n}}\right) &= \frac{1}{n}\sum_{i=1}^{n}\gamma_i^2\mathrm{Var}(x_j) + \frac{2}{n}\sum_{0<j-i<m}\gamma_i\gamma_j\mathrm{Cov}(x_i,x_j)\\
&\to p\left\{c^2\eta_1(1-\eta_1) + 2cd\eta_1(1-\eta_2) + d^2\eta_2(1-\eta_2)\right\}\\
&\quad \times\left(1-p+2\sum_{w=1}^{m}\left(\boldsymbol{P}_{11}^{w}-p\right)\right).
\end{aligned} \tag{2.18}$$

It is clear that the asymptotic variance of $S_{t_1,t_2}/\sqrt{n}$ is finite. The proof of the following lemma uses this fact to show that the asymptotic joint distribution of S_{t_1} and S_{t_2} is bivariate

normal, by relying on Result 4.2 from Johnson and Wichern [18], which is restated below for convenience.

Theorem 2.4.2 *Let $\boldsymbol{X}$ denote a p-dimensional vector of random variables and $\boldsymbol{a}$ a vector of constants. If $\boldsymbol{a}'\boldsymbol{X}$ is distributed as $N(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$, then $\boldsymbol{X}$ must be $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

Lemma 2.4.3 *Suppose the sequence $\{x_t\}_{t=1}^n$ is m-dependent and the value of m is known. The asymptotic joint distribution of T_{t_1} and T_{t_2} is bivariate normal.*

Proof First, the asymptotic joint normality of S_{t_1} and S_{t_2} must be shown. By Theorem 2.4.2, it suffices to show that for any constants $c, d \in \mathbb{R}$, the linear combination $S_{t_1,t_2} = cS_{t_1} + dS_{t_2}$ is normally distributed. Define S_{t_1,t_2} as in equation (2.14), then the random variables $\gamma_j x_j$ that make up S_{t_1,t_2} are uniformly bounded by $|c + d|$.

The above calculations show that the limiting variance of $S_{t_1,t_2}/\sqrt{n}$ is finite and equal to the final expression in equation (2.18). Therefore, $\mathrm{Var}(S_{t_1,t_2})/n^{2/3} \to \infty$. By Theorem 1.4.5:

$$\frac{cS_{t_1} + dS_{t_2}}{\sqrt{\mathrm{Var}(cS_{t_1} + dS_{t_2})}} = \frac{S_{t_1,t_2}}{\sqrt{\mathrm{Var}(S_{t_1,t_2})}} \xrightarrow{D} \mathrm{N}(0,1). \tag{2.19}$$

To conclude the proof, the linear combination $cT_{t_1} + dT_{t_1}$ must be shown to be normally distributed:

$$\begin{aligned} cT_{t_1} + dT_{t_2} &= c\frac{\mathrm{DCUSUM}_{t_1}}{\hat{\sigma}_{D,t_1}} + d\frac{\mathrm{DCUSUM}_{t_2}}{\hat{\sigma}_{D,t_2}} \\ &= c\frac{S_{t_1}}{\sqrt{\widehat{\mathrm{Var}}(S_{t_1})}} + d\frac{S_{t_2}}{\sqrt{\widehat{\mathrm{Var}}(S_{t_2})}} \\ &= c'S_{t_1} + d'S_{t_2}. \end{aligned}$$

By equation (2.19), $cT_{t_1} + dT_{t_2}$ converges in distribution to a normal density. Therefore, T_{t_1} and T_{t_2} are bivariate normal. ∎

The mean vector for the joint distribution of T_{t_1} and T_{t_1+1} is clearly $(0, 0)$ because both T_{t_1} and T_{t_1+1} have zero expectation. The diagonal entries in the covariance matrix are 1, and

the approximate covariance σ_{t_1,t_1+1} can be calculated using equation (2.11) by substituting any consistent estimator $\hat{\sigma}_{t_1,t_1+1}$. To summarize:

$$\begin{pmatrix} T_{t_1} \\ T_{t_1+1} \end{pmatrix} \to \mathrm{N}(\boldsymbol{\mu}, \Sigma), \text{ where } \boldsymbol{\mu} = (0,0) \text{ and } \Sigma \approx \begin{pmatrix} 1 & \hat{\sigma}_{t_1,t_1+1} \\ \hat{\sigma}_{t_1,t_1+1} & 1 \end{pmatrix}.$$

Combining the summary above with Lemmas 2.4.1 and 2.4.3, the Worsley upper bound (2.13) can be estimated.

Chapter 3

Dependent Likelihood Ratio Test

A different approach to handle the m-dependent sequence $y = \{x_t\}_{t=1}^n$ is a generalization of the likelihood ratio test discussed in Section 1.2.3. Instead of maximizing the full likelihood function, a modified likelihood function proposed by Billingsley [2] and implemented in the change point detection setting by Krauth [20, 21] is used. The asymptotic distribution of the test statistic G_t^2 for fixed time t is found and an approximate asymptotic distribution for $G_{\max}^2$, the maximum of G_t^2 over the range of permissible values of t, is proposed. Due to the lack of a known asymptotic distribution for the dependent likelihood ratio test (DLRT) statistic, a bootstrap procedure is proposed to approximate p-values.

3.1 Modified Likelihood Function and MLEs

The one step Markov dependence assumption emphasizes transitions between consecutive random variables in the sequence. To record these values, notation from Krauth [20] is introduced. Define $y^{(t)} = \{x_j\}_{j=1}^t$ to be the sequence y truncated at index t. Let n_{uv}^t denote the number of times the truncated sequence $y^{(t)}$ has transitioned from state u to state v,

that is:

$$n_{11}^t = \sum_{j=2}^{t} x_{j-1}x_j, \quad n_{00}^t = \sum_{j=2}^{t}(1-x_{j-1})(1-x_j),$$
$$n_{10}^t = \sum_{j=2}^{t} x_{j-1}(1-x_j), \quad n_{01}^t = \sum_{j=2}^{t}(1-x_{j-1})x_j. \tag{3.1}$$

The likelihood ratio statistic requires both the likelihood functions under the null and alternative hypotheses. These functions are stated below and are described in more detail in Devore [7] and Krauth [20], respectively. The alternative likelihood function given below is for a fixed time t:

$$\begin{aligned}
L_{H_0} =& p^{x_1}(1-p)^{1-x_1}L_{H_0}^* \text{ where } L_{H_0}^* = P_{00}^{n_{00}}P_{01}^{n_{01}}P_{10}^{n_{10}}P_{11}^{n_{11}},\\
L_{H_a} =& p^{x_1}(1-p)^{1-x_1}P_{00}^{n_{00}^t}(1)P_{01}^{n_{01}^t}(1)P_{10}^{n_{10}^t}(1)P_{11}^{n_{11}^t}(1)\\
&\times P_{11}(t)^{x_t x_{t+1}}P_{10}(t)^{x_t(1-x_{t+1})}P_{01}(t)^{(1-x_t)x_{t+1}}P_{00}(t)^{(1-x_t)(1-x_{t+1})}\\
&\times P_{11}(2)^{n_{11}^n - n_{11}^t - x_t x_{t+1}}P_{10}(2)^{n_{10}^n - n_{10}^t - x_t(1-x_{t+1})}\\
&\times P_{01}(2)^{n_{01}^n - n_{01}^t - (1-x_t)x_{t+1}}P_{00}(2)^{n_{00}^n - n_{00}^t - (1-x_t)(1-x_{t+1})}.
\end{aligned}$$

The initial term x_1 and the term x_t lead to complications in maximizing the full likelihood functions. While direct maximization is not mathematically impossible, the complexity of the solution is unreasonable for practical use. Instead, a modified likelihood function proposed by Billingsley [2] and implemented in the change point detection setting by Krauth [20, 21] is used. The initial term x_1 and the term x_t are ignored in the modified likelihood functions given below, denoted with $*$. These can be maximized in the usual way by taking derivatives.

$$\begin{aligned}
L_{H_0}^* =& P_{00}^{n_{00}^n}P_{01}^{n_{01}^n}P_{10}^{n_{10}^n}P_{11}^{n_{11}^n},\\
L_{H_a}^* =& P_{00}^{n_{00}^t}(1)P_{01}^{n_{01}^t}(1)P_{10}^{n_{10}^t}(1)P_{11}^{n_{11}^t}(1)\\
&\times P_{11}(2)^{n_{11}^n - n_{11}^t - x_t x_{t+1}}P_{10}(2)^{n_{10}^n - n_{10}^t - x_t(1-x_{t+1})}
\end{aligned}$$

$$\times P_{01}(2)^{n_{01}^n - n_{01}^t - (1-x_t)x_{t+1}} P_{00}(2)^{n_{00}^n - n_{00}^t - (1-x_t)(1-x_{t+1})}. \tag{3.2}$$

In the system of equations (1.21) resulting from the one step Markov dependence assumption, it is assumed that the value of p is given. Alternatively, the system (1.21) can be viewed as a system of equations with two free variables, p and P_{11}. If instead, it is assumed that p is unknown, the free variables can be thought of as P_{00} and P_{11}. The modified maximum likelihood estimates are derived for P_{00} and P_{11} from equations (3.2) for a fixed time t:

$$\begin{aligned}
\hat{P}_{11} &= \frac{n_{11}^n}{n_{11}^n + n_{10}^n}, \quad \hat{P}_{00} = \frac{n_{00}^n}{n_{00}^n + n_{01}^n}, \\
\hat{P}_{11}(1) &= \frac{n_{11}^t}{n_{11}^t + n_{10}^t}, \quad \hat{P}_{00}(1) = \frac{n_{00}^t}{n_{00}^t + n_{01}^t}, \\
\hat{P}_{11}(2) &= \frac{n_{11}^n - n_{11}^t - x_t x_{t+1}}{n_{11}^n - n_{11}^t - x_t x_{t+1} + n_{10}^n - n_{10}^t - x_t(1-x_{t+1})}, \\
\hat{P}_{00}(2) &= \frac{n_{00}^n - n_{00}^t - (1-x_t)(1-x_{t+1})}{n_{00}^n - n_{00}^t - (1-x_t)(1-x_{t+1}) + n_{01}^n - n_{01}^t - (1-x_t)x_{t+1}}.
\end{aligned} \tag{3.3}$$

With these estimates in hand, the MLEs for p, p_1, and p_2 are found by substitution of equation (3.7) into the system (1.21):

$$\begin{aligned}
\hat{p} &= \frac{1 - \hat{P}_{00}}{2 - \hat{P}_{00} - \hat{P}_{11}}, \\
\hat{p}(1) &= \frac{1 - \hat{P}_{00}(1)}{2 - \hat{P}_{00}(1) - \hat{P}_{11}(1)}, \\
\hat{p}(2) &= \frac{1 - \hat{P}_{00}(2)}{2 - \hat{P}_{00}(2) - \hat{P}_{11}(2)}.
\end{aligned} \tag{3.4}$$

The value of τ is unknown and must be estimated. There are $n-2$ possible time points for the location of τ, which leads to $n-2$ possible values for the maximum likelihood estimate under the alternative hypothesis. The maximum alternative likelihood is recorded for each possible value of $t = 2, \ldots, n-1$. The global maximum value is taken to be $\max_{2 \le t \le n-1} L_{H_a}^*$.

This yields the modified likelihood ratio statistic:

$$\lambda^* = \frac{L^*_{H_0}}{\max\limits_{2\le t\le n-1} L^*_{H_a}}.$$

From equation (3.2), it is clear that $L^*_{H_0}$ is comprised of exactly one more term than $L^*_{H_a}$. All of the parameters fall in the interval $(0,1)$ forcing the strict inequality $L^*_{H_0} < L^*_{H_a}$, even in the case when H_0 is true. This issue is addressed by removing the transition from t to $t+1$ from the null likelihood function, for each value of t. The updated modified null likelihood function and likelihood ratio are:

$$L^{**}_{H_0} = P_{11}^{n^n_{11}-x_t x_{t+1}} P_{10}^{n^n_{10}-x_t(1-x_{t+1})} P_{01}^{n^n_{01}-(1-x_t)x_{t+1}} P_{00}^{n^n_{00}-(1-x_t)(1-x_{t+1})} \text{ and } \lambda^{**} = \max_{2\le t\le n-1} \frac{L^{**}_{H_0}}{L^*_{H_a}}.$$

The resulting MLEs for $P_{00}, P_{11},$ and p are given below, with $\hat{\tau}$ defined in the following paragraph:

$$\begin{aligned}
\hat{P}_{00} &= \frac{n^n_{00} - (1-x_{\hat{\tau}})(1-x_{\hat{\tau}+1})}{n^n_{00} - (1-x_{\hat{\tau}})(1-x_{\hat{\tau}+1}) + n^n_{01} - (1-x_{\hat{\tau}})x_{\hat{\tau}+1}},\\
\hat{P}_{11} &= \frac{n^n_{11} - x_{\hat{\tau}}x_{\hat{\tau}+1}}{n^n_{11} - x_{\hat{\tau}}x_{\hat{\tau}+1} + n^n_{10} - x_{\hat{\tau}}(1-x_{\hat{\tau}+1})},\\
\hat{p} &= \frac{1-\hat{P}_{00}}{2-\hat{P}_{00}-\hat{P}_{11}}.
\end{aligned} \tag{3.5}$$

The likelihood functions take on values that are extremely close to zero, making computations very difficult for a computer. Instead, the G^2_t statistic is used, where:

$$G^2_t = 2(\log L^*_{H_a} - \log L^{**}_{H_0}). \tag{3.6}$$

Large values of G^2_t will be evidence that a change point exists at the value $\hat{\tau} = \arg\max\limits_t G^2_t$. By construction, the permissible range of values of t where G^2_t is well defined depends on the sequence y. While restricting to the range of values $nl \le t \le nh$ similar to DCUSUM is one solution, in this dissertation, the values of t are found on a case by case basis. The

asymptotic distribution of G_t^2 for fixed t and approximate asymptotic distribution of $G_{\max}^2$ are discussed in the next section.

3.2 Asymptotic Distribution of G_t^2

The goal of this section is to determine the asymptotic distribution of the test statistic G_t^2 for a fixed time $1 < t < n$. The definition of the test statistic is given by equation (3.6).

For fixed t, the contribution of the transition from x_t to x_{t+1} to the MLEs of $P_{00}, P_{11}, P_{00}(2)$, and $P_{11}(2)$ is only a single time point and will be lost in the limit. Therefore, those terms in equations (3.7) and (3.5) are ignored in the large sample MLEs, which are given below:

$$\begin{aligned}
\hat{P}_{11} &= \frac{n_{11}^n}{n_{11}^n + n_{10}^n}, \quad \hat{P}_{00} = \frac{n_{00}^n}{n_{00}^n + n_{01}^n}, \\
\hat{P}_{11}(1) &= \frac{n_{11}^t}{n_{11}^t + n_{10}^t}, \quad \hat{P}_{00}(1) = \frac{n_{00}^t}{n_{00}^t + n_{01}^t}, \\
\hat{P}_{11}(2) &= \frac{n_{11}^n - n_{11}^t}{n_{11}^n - n_{11}^t + n_{10}^n - n_{10}^t}, \quad \hat{P}_{00}(2) = \frac{n_{00}^n - n_{00}^t}{n_{00}^n - n_{00}^t + n_{01}^n - n_{01}^t}.
\end{aligned} \tag{3.7}$$

It is well known that under certain conditions, likelihood ratio statistics follow a χ^2 asymptotic distribution. The main result from Wilks [38] gives criteria for this to occur. For the purposes of this dissertation those criteria are that the large sample joint distribution of the MLEs $\hat{P}_{uv}(1)$ and $\hat{P}_{uv}(2)$ belongs to an exponential family.

Before proving several lemmas that lead to the main result, some notation is introduced. Let $t^* = n - t$ and $n_{uv}^{t^*} = n_{uv}^n - n_{uv}^t$. For fixed t, the statistics $\hat{P}_{uv}(1)$ and $\hat{P}_{uv}(2)$ from equations (3.7) can be written as:

$$\begin{aligned}
\hat{P}_{uv}(1) &= \frac{n_{uv}^t/(t-1)}{(n_{uv}^t + n_{uv'}^t)/(t-1)} = \frac{\bar{n}_{uv}^t}{(1/(t-1))\sum_{j=2}^t 1_{\{x_j=u\}}}, \\
\hat{P}_{uv}(2) &= \frac{n_{uv}^{t^*}/(t^*-1)}{(n_{uv}^{t^*} + n_{uv'}^{t^*})/(t^*-1)} = \frac{\bar{n}_{uv}^{t^*}}{(1/(t^*-1))\sum_{j=t+2}^n 1_{\{x_j=u\}}}.
\end{aligned} \tag{3.8}$$

The claim is that the random vector:

$$\hat{\boldsymbol{P}} = \left(\hat{P}_{11}(1)\ \hat{P}_{00}(1)\ \hat{P}_{11}(2)\ \hat{P}_{00}(2)\right)',$$

has an asymptotic multivariate normal distribution with mean vector:

$$\boldsymbol{\mu} = (P_{11}(1)\ P_{00}(1)\ P_{11}(2)\ P_{00}(2))',$$

and finite covariance matrix Σ.

First, it is shown that the denominators in equations (3.8) converge almost surely to constants. The law of large numbers for Markov chains given in Durrett [8], Theorem 6.6.1, states:

Theorem 3.2.1 *Suppose u is recurrent and define $E_u R_u$ to be the expected amount of time until x_t returns to state u. For any state u, as $n \to \infty$:*

$$\frac{\sum_{j=1}^{n} 1_{\{x_j=u\}}}{n} \to \frac{1}{E_u R_u} 1_{\{T_u<\infty\}} \quad P_x - a.s.$$

The value of the limit is given in Durrett [8], Theorem 6.5.5, which states:

Theorem 3.2.2 *If a Markov chain is irreducible and has stationary distribution π, then $\pi(u) = 1/E_u R_u$.*

Combining this theorem with the stationary distribution given in Lemma 1.4.2 yields the following convergence result:

Lemma 3.2.3 *The values $(1/(t-1))\sum_{j=2}^{t} 1_{\{x_j=u\}}$ and $(1/(t^*-1))\sum_{j=t+2}^{n} 1_{\{x_j=u\}}$ converge almost surely to constants which depend on the value of u.*

Proof An application of Theorem 3.2.1 followed by an application of Theorem 3.2.2 to the

denominators of equations (3.8) will prove the result.

$$\frac{1}{t-1}\sum_{j=2}^{t} 1_{\{x_j=u\}} \to \frac{1}{E_u(1)R_u(1)} 1_{\{R_u(1)<\infty\}} = p(1) \text{ or } (1-p(1)),$$
$$\frac{1}{t^*-1}\sum_{j=t+2}^{n} 1_{\{x_j=u\}} \to \frac{1}{E_u(2)R_u(2)} 1_{\{R_u(2)<\infty\}} = p(2) \text{ or } (1-p(2)).$$

The right hand sides of both equations are constants dependent on the value of $u = 0$ or 1.

∎

The next step in proving the main result of this section is to show that the asymptotic distributions of the numerators in equations (3.8) are normal. For the remainder of this section, only the pre change case where $u = v = 1$ (that is, n_{11}^t) is considered, as the other three cases are similar.

Define $z_i = x_{i-1}x_i$. Under the alternative hypothesis, z_i is a Bernoulli random variable with success probability:

$$\Pr(z_i = 1) = \begin{cases} p(1)P_{11}(1) & \text{if } 2 \le i \le t, \\ p(2)P_{11}(2) & \text{if } t+2 \le i \le n, \end{cases}$$

and covariance:

$$\text{Cov}(z_i, z_j) = \begin{cases} p(1)P_{11}^2(1)\left(\boldsymbol{P}_{11}^{j-i-1}(1) - p(1)\right) & \text{if } 1 \le i < j \le t, \\ 0 & \text{if } 1 \le i \le t < j \le n, \\ p(2)P_{11}^2(2)\left(\boldsymbol{P}_{11}^{j-i-1}(2) - p(2)\right) & \text{if } t_2 \le i < j \le n. \end{cases}$$

For consecutive terms, the covariance reduces to $\text{Cov}(z_i, z_{i+1}) = p(1)(1-p(1))P_{11}^2(1)$, $p(2)(1-p(2))P_{11}^2(2)$, or 0.

Similar to Lemma 1.4.3 as $n \to \infty$, $j - i \to \infty$ for all but finitely many terms. Therefore, for large n, $\boldsymbol{P}_{11}^{j-i-1}(1) \approx p(1)$ and $\boldsymbol{P}_{11}^{j-i-1}(2) \approx p(2)$. Using a similar argument as Lemma 1.4.4, the sequence of correlated Bernoulli trials $\{z_i\}_{i=2}^t$ that make up n_{11}^t may be considered

to be approximately m-dependent.

In order to apply Theorem 1.4.5 to obtain the distribution of n_{11}^t, the variance of n_{11}^t, say $V_{n_{11}^t}$, must be calculated:

$$\begin{aligned} V_{n_{11}^t} = \text{Var}\left(n_{11}^t\right) = \text{Var}\left(\sum_{i=2}^{t} z_i\right) &= \sum_{i=2}^{t} \text{Var}(z_i) + 2 \sum_{2 \le i < j \le t} \text{Cov}(z_i, z_j) \\ &= (t-1)p(1)P_{11}(1)\left(1 - p(1)P_{11}(1)\right) \\ &\quad + 2 \sum_{0 < j-i \le m}^{t} p(1)P_{11}^2(1)\left(\boldsymbol{P}_{11}^{j-i-1}(1) - p(1)\right). \end{aligned} \tag{3.9}$$

It is clear from equation (3.9) that the variance of $\bar{n}_{11}^t$ is defined as:

$$V_{\bar{n}_{11}^t} = V_{n_{11}^t}/(t-1)^2. \tag{3.10}$$

Lemma 3.2.4 *Suppose the sequence $\{z_i\}_{i=2}^t$ is m-dependent and the value of m is known. The statistic $\bar{n}_{11}^t$ is asymptotically normal with mean $p(1)P_{11}(1)$ and finite variance.*

Proof The random variables that make up n_{11}^t are uniformly bounded by 1. From equation (3.9), it is clear that $V_{n_{11}^t}$ is of order t. Therefore, $V_{n_{11}^t}/(t-1)^{2/3} \to \infty$. By Theorem 1.4.5:

$$\frac{\bar{n}_{11}^t - p(1)P_{11}(1)}{\sqrt{V_{\bar{n}_{11}^t}}} = \frac{n_{11}^t - (t-1)p(1)P_{11}(1)}{\sqrt{V_{n_{11}^t}}} \xrightarrow{d} \text{N}(0,1).$$

∎

Combining Lemmas 3.2.3 and 3.2.4 yields the first main result of this section.

Theorem 3.2.5 *Suppose the sequence $\{z_i\}_{i=2}^t$ is m-dependent and the value of m is known. Define n_{uv}^t as in equation (3.1), $\hat{P}_{11}(1)$ as in equation (3.7), and $V_{n_{11}^t}$ as in equation (3.9). The statistic $\hat{P}_{11}(1)$ is asymptotically normal with mean $P_{11}(1)$ and finite variance.*

Proof Define $V_{11} = \lim_{n \to \infty} V_{n_{11}^t}/(t-1) = p(1)P_{11}(1)\left(1 - p(1)P_{11}(1)\right) + C_{11}$ where C_{11} is the

limit of the second term in equation (3.9). By Slutsky's Theorem:

$$\begin{aligned}
\sqrt{t-1}\left(\hat{P}_{11}(1)-P_{11}(1)\right) &= \sqrt{t-1}\left(\frac{\bar{n}_{11}^t}{(n_{11}^t+n_{10}^t)/(t-1)}-P_{11}(1)\right)\\
&= \sqrt{t-1}\left(\frac{\bar{n}_{11}^t}{(n_{11}^t+n_{10}^t)/(t-1)}-\frac{p(1)P_{11}(1)}{(n_{11}^t+n_{10}^t)/(t-1)}\right.\\
&\qquad\left.+\frac{p(1)P_{11}(1)}{(n_{11}^t+n_{10}^t)/(t-1)}-P_{11}(1)\right)\\
&= \sqrt{t-1}\left(\bar{n}_{11}^t-p(1)P_{11}(1)\right)\left(\frac{1}{t-1}\sum_{i=2}^{t}1_{\{x_i=1\}}\right)^{-1}\\
&\qquad+\sqrt{t-1}\left(\frac{p(1)}{(n_{11}^t+n_{10}^t)/(t-1)}-1\right)P_{11}(1)\\
&\to \frac{1}{p(1)}N\left(0,V_{11}\right)=N\left(0,\frac{V_{11}}{p^2(1)}\right).
\end{aligned}$$

∎

Notice that the variation of $\hat{P}_{11}(1)$ tends to 0 as t tends to infinity. This implies that the estimate $\hat{P}_{11}(1)$ of $P_{11}(1)$ is consistent.

Corollary 3.2.6 *The MLE $\hat{P}_{11}(1)$ is a consistent estimator of $P_{11}(1)$.*

A similar argument shows that the centered and scaled version of the estimator $\hat{P}_{00}(1)$ is asymptotically normal and that $\hat{P}_{00}(1)$ is a consistent estimator of $P_{00}(1)$. The asymptotic distributions of $\hat{P}_{11}(2)$ and $\hat{P}_{00}(2)$ may be found by substituting (2) for (1) in the arguments above.

The next step is to determine the joint distribution of $\hat{P}_{11}(1)$ and $\hat{P}_{00}(1)$.

Theorem 3.2.7 *$\hat{P}_{11}(1)$ and $\hat{P}_{00}(1)$ are asymptotically bivariate normal.*

Proof The goal is to show that any linear combination $a\hat{P}_{11}(1)+b\hat{P}_{00}(1)$ is asymptotically

normal for any constants a and b. This can be re written as:

$$\begin{aligned}
a\hat{P}_{11}(1) + b\hat{P}_{00}(1) &= \frac{an_{11}^t}{n_{11}^t + n_{10}^t} + \frac{bn_{00}^t}{n_{00}^t + n_{01}^t} \\
&= \frac{a\sum_{j=2}^t x_{j-1}x_j}{\sum_{l=2}^t x_{l-1}} + \frac{b\sum_{j=2}^t (1-x_{j-1})(1-x_j)}{\sum_{l=2}^t (1-x_{l-1})} \\
&= \left(\frac{1}{\sum_{l=2}^t x_{l-1}}\right)\left(\frac{1}{\sum_{l=2}^t (1-x_{l-1})}\right) \\
&\quad \times \left[a\left(\sum_{j=2}^t x_{j-1}x_j\right)\left(\sum_{l=2}^t (1-x_{l-1})\right)\right. \\
&\quad \left. + b\left(\sum_{j=2}^t (1-x_{j-1})(1-x_j)\right)\left(\sum_{l=2}^t x_{l-1}\right)\right] \\
&= \left(\frac{1}{\sum_{l=2}^t x_{l-1}}\right)\left(\frac{1}{\sum_{l=2}^t (1-x_{l-1})}\right)(t-1)\left[a\sum_{j=2}^t x_{j-1}x_j\right. \\
&\quad \left. + \left(b\sum_{j=2}^t (1-x_{j-1})(1-x_j) - a\sum_{j=2}^t x_{j-1}x_j\right)\left(\frac{\sum_{l=2}^t x_{l-1}}{t-1}\right)\right] \\
&:= \left(\frac{1}{\sum_{l=2}^t x_{l-1}}\right)\left(\frac{1}{\sum_{l=2}^t (1-x_{l-1})}\right)(t-1)\sum_{j=2}^t c_j,
\end{aligned}$$

where:

$$a' = a\left(1 - \frac{\sum_{l=2}^t x_{l-1}}{t-1}\right),\ b' = b\left(\frac{\sum_{l=2}^t x_{l-1}}{t-1}\right),\ \text{and } c_j = \begin{cases} a' & \text{if } x_{j-1} = x_j = 1, \\ b' & \text{if } x_{j-1} = x_j = 0, \\ 0 & \text{else.} \end{cases}$$

The sequence $\{c_j\}_{j=2}^t$ is approximately m-dependent. Two of the cases are shown below, as the others are similar.

Fix m as determined by the transition matrix $\boldsymbol{P}(1)$ and, without loss of generality, suppose $i + m < j$. Then:

$$\begin{aligned}
\Pr(c_j = a' \cap c_i = a') &= \Pr(x_j = 1|x_{j-1} = 1)\Pr(x_{j-1} = 1|x_i = 1) \\
&\quad \times \Pr(x_i = 1|x_{i-1} = 1)\Pr(x_{i-1} = 1)
\end{aligned}$$

$$
\begin{aligned}
&= P_{11}(1)\boldsymbol{P}_{11}^{j-i-1}(1)P_{11}(1)p(1) \\
&\approx [P_{11}(1)p(1)][P_{11}(1)p(1)] \\
&= \Pr(c_j = a')\Pr(c_i = a'). \\
\Pr(c_j = a' \cap c_i = b') &= \Pr(x_j = 1|x_{j-1} = 1)\Pr(x_{j-1} = 1|x_i = 0) \\
&\qquad \times \Pr(x_i = 0|x_{i-1} = 0)\Pr(x_{i-1} = 0) \\
&= P_{11}(1)\boldsymbol{P}_{01}^{j-i-1}(1)P_{00}(1)(1 - p(1)) \\
&\approx [P_{11}(1)p(1)][P_{00}(1)(1 - p(1))] \\
&= \Pr(c_j = a')\Pr(c_i = b').
\end{aligned}
$$

Note that the random variables c_j are uniformly bounded by $\max\{|a|, |b|\}$. The variance of $\sum c_j$ is:

$$
\begin{aligned}
\operatorname{Var}\left(\sum_{j=2}^{t} c_j\right) &= \sum_{j=2}^{t} \operatorname{Var}(c_j) + 2\sum_{i<j} \operatorname{Cov}(c_i, c_j) \\
&= (t-1)\left((a')^2 P_{11}(1)p(1) + (b')^2 P_{00}(1)(1 - p(1))\right. \\
&\quad -[a'P_{11}(1)p(1) + b'P_{00}(1)(1 - p(1))]^2\big) \\
&\quad + 2\sum_{0<i-j\le m} \left[(a')^2 P_{11}^2(1)p(1)\boldsymbol{P}_{11}^{j-i-1}(1) + (b')^2 P_{00}^2(1)(1 - p(1))\boldsymbol{P}_{00}^{j-i-1}(1)\right. \\
&\qquad + a'b'P_{00}(1)P_{11}(1)\left(\boldsymbol{P}_{01}^{j-i-1}(1)(1 - p(1)) + \boldsymbol{P}_{10}^{j-i-1}(1)p(1)\right) \\
&\qquad \left. - (a'P_{11}(1)p(1) + b'P_{00}(1)(1 - p(1)))^2\right].
\end{aligned}
$$

It is clear that $\operatorname{Var}\left(\sum c_j\right)$ is of order t, so $\operatorname{Var}\left(\sum c_j\right)/(t-1)^{2/3} \to \infty$. As t tends to infinity, $a' \to a(1 - p(1))$ and $b' \to bp(1)$. Applying Theorem 1.4.5 gives the convergence of $\sum c_j$:

$$
\begin{aligned}
&\frac{\frac{\sum_{j=2}^{t} c_j}{t-1} - p(1)(1 - p(1))\left(aP_{11}(1) + bP_{00}(1)\right))}{\operatorname{Var}\left(\sum_{j=2}^{t} c_j\right)\Big/(t-1)} \\
&= \frac{\sum_{j=2}^{t} c_j - (t-1)p(1)(1 - p(1))\left(aP_{11}(1) + bP_{00}(1)\right))}{\operatorname{Var}\left(\sum_{j=2}^{t} c_j\right)} \to N(0, 1).
\end{aligned}
$$

Therefore, $\left(\sum_{j=2}^{t} c_j\right) \Big/ \sqrt{t-1}$ is approximately normal with mean $p(1)(1-p(1))(aP_{11}(1) + bP_{00}(1))$ and finite variance, say σ_c^2. Combining this result with Slutsky's Theorem and Lemma 3.2.3 gives the asymptotic normality of $\hat{P}_{11}(1)$ and $\hat{P}_{00}(1)$.

Define $\mu_{ab} = (aP_{11}(1) + bP_{00}(1))$. Then:

$$\begin{aligned}
\sqrt{t-1}\left(a\hat{P}_{11}(1) + b\hat{P}_{00}(1) - \mu_{ab}\right) &= \sqrt{t-1}\left(a\hat{P}_{11}(1) + b\hat{P}_{00}(1)\right. \\
&\qquad - \frac{(t-1)^2 p(1)(1-p(1))\mu_{ab}}{\left(\sum_{j=2}^{t} x_{j-1}\right)\left(\sum_{j=2}^{t}(1-x_{j-1})\right)} \\
&\qquad \left. + \frac{(t-1)^2 p(1)(1-p(1))\mu_{ab}}{\left(\sum_{j=2}^{t} x_{j-1}\right)\left(\sum_{j=2}^{t}(1-x_{j-1})\right)} - \mu_{ab}\right) \\
&= \sqrt{t-1}\left(\frac{\sum_{j=2}^{t} c_j}{t-1} - p(1)(1-p(1))\mu_{ab}\right) \\
&\qquad \times \left(\frac{(t-1)^2}{\left(\sum_{j=2}^{t} x_{j-1}\right)\left(\sum_{j=2}^{t}(1-x_{j-1})\right)}\right) \\
&\qquad + \sqrt{t-1}\left(\frac{(t-1)^2 p(1)(1-p(1))}{\left(\sum_{j=2}^{t} x_{j-1}\right)\left(\sum_{j=2}^{t}(1-x_{j-1})\right)} - 1\right)\mu_{ab} \\
&\xrightarrow{d} \frac{1}{p(1)(1-p(1))}\mathrm{N}\left(0, \sigma_c^2\right) \\
&= \mathrm{N}\left(0, \frac{\sigma_c^2}{p^2(1)(1-p(1))^2}\right).
\end{aligned}$$

The choice of a and b was arbitrary, implying that any linear combination of $P_{11}(1)$ and $P_{00}(1)$ is asymptotically normal. By Theorem 2.4.2, $P_{11}(1)$ and $P_{00}(1)$ have a bivariate normal joint asymptotic distribution. ∎

The asymptotic bivariate normality of $\hat{P}_{11}(2)$ and $\hat{P}_{00}(2)$ can be shown by interchanging (1) for (2) and t for t^*.

To conclude this section, recall that the alternative model assumes that there is an abrupt change at the fixed time point t. Therefore, the random variables x_j, $1 \leq j \leq t$, are

independent of those values x_j, $t < j \leq n$. Theorem 3.2.7 implies that $\hat{P}_{11}(1)$ and $\hat{P}_{00}(1)$ are asymptotically bivariate normal with mean vector $\boldsymbol{\mu}(1) = (P_{11}(1)\ P_{00}(1))$ and finite covariance matrix, say $\Sigma(1)$, and that the statistics $\hat{P}_{11}(2)$ and $\hat{P}_{00}(2)$ are asymptotically bivariate normal with mean vector $\boldsymbol{\mu}(2) = (P_{11}(2)\ P_{00}(2))$ and finite covariance matrix, say $\Sigma(2)$. These two bivariate normal random vectors are independent by the assumptions of the model, so by Result 4.5 (c) in Johnson and Wichern [18], the joint random vector $\hat{\boldsymbol{P}} = \left(\hat{P}_{11}(1)\ \hat{P}_{00}(1)\ \hat{P}_{11}(2)\ \hat{P}_{00}(2)\right)'$ is asymptotically multivariate normal.

The asymptotic distribution of G_t^2 is stated in the following theorem.

Theorem 3.2.8 *For a fixed time point t, where $1 < t < n$, the distribution of G_t^2 is asymptotically χ_2^2.*

Proof The joint distribution of the MLEs of G_t^2 is asymptotically multivariate normal, which belongs to an exponential family. By the Theorem in Wilks [38], the asymptotic distribution of G_t^2, except for terms of order $1/\sqrt{n}$, is χ_{h-m}^2. For each fixed time t, $h = 4$, as there are four parameters in the alternative model, and $m = 2$, as there are two parameters in the null model. Therefore, $G_t^2 \xrightarrow{d} \chi_2^2$. ∎

While the distribution of G_t^2 is known for fixed t, the covariance structure of $G_{t_1}^2$ and $G_{t_2}^2$ for $t_1 \neq t_2$ is very complex. Because of this, it is an open problem to determine the asymptotic distribution of $G_{\max}^2 = \max_t G_t^2$.

Hinkley [12] and Feder [9] discuss the distribution of this type of statistic in the change point problem for regression models. Specifically, the hypotheses test for a single change point in the simple linear regression fit of a single sequence of independent observations. The empirical conclusion is that $G_{\max}^2 \approx \chi_3^2$. Similar empirical results are shown for the dependent Bernoulli sequence change point problem in Section 4.3.2, but with larger degrees of freedom.

The distribution of the test statistic $G_{\max}^2$ depends on the value of the parameter P_{11}. This can be seen by the simulations in Section 4.3.1, and in particular, the percentile values

in Tables 4.8, 4.9, 4.10, and 4.11. Berger and Boos [1] proposed a method to approximate p-values in similar situations by maximizing the p-value over a confidence set for the parameter P_{11}. This method can be applied by maximizing over the set of permissible values of P_{11} as given in equation (1.22).

3.3 Bootstrap p-value Approximation

The lack of a known asymptotic distribution for the maximum likelihood ratio statistic causes complications in formal hypothesis testing using the $G^2_{\max}$ statistic. A bootstrap approximation is introduced to approximate the distribution of $G^2_{\max}$ under the hypothesis of no change and estimate p-values for change point detection. The algorithm is described first, followed by an example and proofs.

3.3.1 The Bootstrap Algorithm

The one step Markov dependence assumption on the sequence of Bernoulli trials $y = \{x_t\}_{t=1}^n$ violates the basic bootstrap requirement that the variables in the sequence are exchangeable. Instead, the assumption creates runs of 0 and 1 to occur in the sequence. It is possible that the runs may be treated as exchangeable components of an m-dependent binary sequence.

Suppose that m is known. The algorithm begins by recording the lengths and values of each run. Define $R_0 := \{r \mid r = \text{length of a run of 0 in the original sequence}\}$ and $R_1 := \{r \mid r = \text{length of a run of 1 in the original sequence}\}$. After the elements of R_0 and R_1 are recorded, it is necessary to extract runs that are independent of one another for use in bootstrap re sampling. For notational purposes, let $r_{i,j}$ be the elements of the set R_j for $j = 0, 1$.

Define $R_0^m \subset R_0$ and $R_1^m \subset R_1$ such that the elements in each of R_0^m and R_1^m are independent. The algorithm to construct R_0^m is described below. To construct R_1^m, interchange the roles of 0 and 1.

Begin by randomly selecting an element $r_{0,0}$ from R_0. This will function as the starting point for the construction of the set of independent elements R_0^m. If there are any elements of R_0 that occur after $r_{0,0}$, select the first run with starting index greater than m plus the ending index of $r_{0,0}$ and call this element $r_{1,0}$. Repeat this process, starting with $r_{1,0}$ and continue until there are no more runs to select.

Next, work backward to determine if there are any elements of R_0 that occur before $r_{0,0}$. If any exist, select the first run with ending index less than the beginning index of $r_{0,0}$ minus m, and call this element $r_{-1,0}$. Repeat this process, starting with $r_{-1,0}$ and continue until there are no more runs to select.

The elements of R_0^m and R_1^m are the elements that will be used to construct the bootstrap sample. Notice that the elements of R_j^m need not be unique.

Once the sets R_0^m and R_1^m have been selected, the bootstrap sequence is constructed in the following way. If the sequence begins with a run of 0, randomly select an element of R_0^m to start the sequence. To complete the sequence, randomly select elements alternating between sets R_1^m and R_0^m until the re sampled sequence has length at least n, and truncate the sequence if it has length greater than n. If the sequence begins with a run of 1, interchange the roles of 0 and 1 in the process. Repeat this process B times using the same starting point $r_{0,0}$ and $r_{1,1}$ each time.

There are three limitations to this bootstrap method. In practice, m is unknown and must be estimated from the data. The estimation process is described in Section 4.1. The sample size necessary to apply the convergence results discussed in Section 3.3.3 is quite large. For small samples, other methods may be more appropriate. The run time of the bootstrap algorithm to generate a p-value is long, making it unreasonable to use for samples larger than $n = 500$.

Table 3.1: Values of R_0^m and R_1^m for various choices of $r_{0,0}$ and $r_{0,1}$

$r_{0,0}$	R_0^m
4	$\{4,3,3\}$
2	$\{2,3\}$
3	$\{3,3,4\}$
3	$\{3,3,4\}$

$r_{0,1}$	R_1^m
3	$\{3,4\}$
2	$\{2,3\}$
4	$\{4,3\}$
3	$\{3,2\}$

3.3.2 Minimal Bootstrap Example

An example of this algorithm is described below. Suppose that $m = 4$ and consider the observed sequence:

$$y = 0\,0\,0\,0\,1\,1\,1\,0\,0\,1\,1\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,1\,1\,1.$$

The sets R_0 and R_1 are:

$$R_0 = \{4,2,3,3\}, \quad R_1 = \{3,2,4,3\}.$$

The sets R_0^m and R_1^m for possible choices of $r_{0,0}$ and $r_{0,1}$ are given in Table 3.1.

After the starting point is selected, fill the sequence of length $n = 24$ by randomly selecting alternate elements from R_0^m and R_1^m and truncate when the re sampled sequence has length $n \geq 24$. Repeat B times.

3.3.3 Bootstrap Justification

Justification of the bootstrap method described in Section 3.3.1 will take several steps. First, the elements of R_0^m and R_1^m are shown to be *iid* with asymptotic geometric distribution and parameters P_{01} and P_{10} respectively. Next, a theorem from Mammen [24] is applied to show that the difference of the bootstrap estimate and the MLE of the mean of a geometric sequence converge in probability to 0. Maximum likelihood estimation of the parameters P_{01}

and P_{10} is justified because the MLE of the mean of a geometric distribution is consistent. Last, the modified likelihood functions of the bootstrap and original sample are shown to be asymptotically equivalent.

Lemma 3.3.1 *Let $y = \{x_t\}_{t=1}^n$ be a sequence of Bernoulli random variables with one step Markov dependence defined by the transition matrix $\boldsymbol{P}$ and define $r_{n,0}$ to be the last run of 0 in the sequence. If $r_{n,0}$ is not the last run of y, the elements of R_0^m are independent and identically distributed, otherwise the elements of $R_0^m \setminus r_{n,0}$ are independent and identically distributed. Furthermore, the distribution of the elements of R_0^m converges to a geometric distribution with parameter P_{01}.*

Proof Let R_0^m be defined as in Section 3.3.1 and $r_{\cdot,0} \in R_0^m$ be any element of R_0^m. By definition, $r_{\cdot,0}$ takes a value in the set $\{1, 2, \ldots, n\}$ and must begin with a 0 entry. Suppose $r_{\cdot,0}$ begins at time $t = t_0$. Let $1 \leq z_{\cdot,0} \leq n - t_0$, then $\Pr(r_{\cdot,0} = z_{\cdot,0})$ is determined by the transition probability P_{01} as shown below:

$$\begin{aligned}
\Pr(r_{\cdot,0} = z_{\cdot,0} \mid x_{t_0} = 0) &= \Pr(x_{z_{\cdot,0}+t_0} = 1 \mid x_{z_{\cdot,0}+t_0-1} = 0) \prod_{t=t_0}^{z_{\cdot,0}+t_0-1} \Pr(x_{t+1} = 0 \mid x_t = 0) \\
&= \Pr(x_{z_{\cdot,0}} = 1 \mid x_{z_{\cdot,0}-1} = 0) \Pr(x_{t+1} = 0 \mid x_t = 0)^{z_{\cdot,0}-1} \\
&= P_{01} P_{00}^{z_{\cdot,0}-1} \\
&= (1 - P_{01})^{z_{\cdot,0}-1} P_{01}. \qquad (3.11)
\end{aligned}$$

As a result of the memoryless property of one step Markov dependence, the final expression is independent of the initial time t_0. When $z_{\cdot,0} = n - t_0 + 1$, there is no P_{01} term in (3.11). Therefore, the *pmf* of $r_{\cdot,0}$ may be written as:

$$f_{r_{\cdot,0}}(z_{\cdot,0}) = \begin{cases} (1 - P_{01})^{z_{\cdot,0}-1} P_{01} & \text{if } z_{\cdot,0} = 1, 2, \ldots, n - t_0, \\ (1 - P_{01})^{n-1} & \text{if } z_{\cdot,0} = n - t_0 + 1, \\ 0 & \text{else.} \end{cases}$$

By construction, the elements of R_0^m are independent.

If y ends with a run of 1, then the case where $z_{\cdot,0} = n - t_0 + 1$ does not occur. In this case, it is clear that the elements of R_0^m are *iid*. If y ends with a run of 0, the distribution of $r_{n,0}$ does not have the extra P_{01} term. This leads to a distribution that is not the same as all other elements of R_0^m. To guarantee that all elements are identically distributed, the last element of R_0 must be excluded from the initial choice of $r_{0,0}$ and from the final set R_0^m.

To show convergence to a geometric distribution, the cdf of $r_{\cdot,0}$ for fixed n and t_0 is:

$$F_{n,r_{\cdot,0}}(z_{\cdot,0}) = \begin{cases} \sum_{i=1}^{\lfloor z_{\cdot,0} \rfloor} (1 - P_{01})^{i-1} P_{01} & \text{if } 1 \leq z_{\cdot,0} \leq n - t_0, \\ 1 & \text{if } z_{\cdot,0} > n - t_0, \\ 0 & \text{else.} \end{cases}$$

Hence:

$$\lim_{n\to\infty} F_{n,r_{\cdot,0}}(z_{\cdot,0}) = \begin{cases} \sum_{i=1}^{\lfloor z_{\cdot,0} \rfloor} (1 - P_{01})^{i-1} P_{01} & \text{if } z_{\cdot,0} \geq 1, \\ 0 & \text{else.} \end{cases}$$

The limit is the cdf of a geometric random variable with parameter P_{01}. ∎

Lemma 3.3.2 *Let $y = \{x_t\}_{t=1}^n$ be a sequence of Bernoulli random variables with one step Markov dependence defined by the transition matrix $\boldsymbol{P}$ and define $r_{n,1}$ to be the last run of 1 in the sequence. If $r_{n,1}$ is not the last run of y, the elements of R_1^m are independent and identically distributed, otherwise the elements of $R_1^m \setminus r_{n,1}$ are independent and identically distributed. Furthermore, the distribution of the elements of R_1^m converges to a geometric distribution with parameter P_{10}.*

Proof Follow the proof of Lemma 3.3.1 and substitute 1 for 0. ∎

One of the main results of this section is to show that the difference of the bootstrap estimate of P_{01} and the MLE of the mean of a geometric sequence converge in probability to

0, and that the MLE of the mean of a geometric sequence is a consistent estimator of P_{01} (or P_{10}). The argument is based on an argument described by Horowitz [16], utilizing a theorem from Mammen [24], and justifies the convergence to zero of the difference of the bootstrap estimate of P_{01} and the MLE of the mean of a geometric distribution in probability.

In order to define consistency in the bootstrap sense as defined by Horowitz [16], some notation is required. Let F_0 be the *cdf* of the random sample $\{x_i | i = 1, \ldots, n\}$, $T_n = T_n(x_1, \ldots, x_n)$ be a statistic, $G_n(\nu, F_0) = \Pr(T_n \leq \nu)$ be the exact, finite-sample *cdf* of T_n, $G_\infty(\cdot, F_0)$ be the asymptotic distribution of T_n, and F_n be the empirical distribution function of the data. Suppose that F_n is a consistent estimator of F_0 and let P_n denote the joint distribution of the sample $\{x_i | i = 1, \ldots, n\}$, then, under other conditions described by Horowitz [16], the bootstrap estimator $G_n(\cdot, F_n)$ is *consistent* if for each $\epsilon > 0$:

$$\lim_{n\to\infty} P_n \left[\sup_{\nu} |G_n(\nu, F_n) - G_\infty(\nu, F_0)| > \epsilon \right] = 0.$$

For convenience, the theorem from Mammen [24] is restated below.

Theorem 3.3.3 *Let $\{x_i \mid i = 1, \ldots, n\}$ be a random sample from a population. For a sequence of functions g_n and sequences of numbers ζ_n and σ_n, define $\bar{g}_n = n^{-1} \sum_{i=1}^n g_n(x_i)$ and $T_n = (\bar{g}_n - \zeta_n)/\sigma_n$. For the bootstrap sample $\{x_i^* \mid i = 1, \ldots, n\}$, define $\bar{g}_n^* = n^{-1} \sum_{i=1}^n g_n(x_i^*)$ and $T_n^* = (\bar{g}_n^* - \bar{g}_n)/\sigma_n$. Let $G_n(\nu) = Pr(T_n \leq \nu)$ and $G_n^*(\nu) = Pr^*(T_n^* \leq \nu)$ where Pr^* is the probability distribution induced by bootstrap sampling. Then $G_n^*(\cdot)$ consistently estimates G_n if and only if $T_n \xrightarrow{d} N(0, 1)$.*

The first lemma below shows that the difference of the bootstrap estimator and the MLE of the mean of a geometric sequence converges in probability to 0. This lemma is specific to the set R_0^m, but can easily be modified for R_1^m.

Lemma 3.3.4 *Let R_0^m be an iid random sample of runs of zero from the original sequence of dependent Bernoulli trials y. Define k_0 to be the total number of runs of 0 necessary to achieve*

a bootstrap sequence of length n. The bootstrap estimate of the sample mean $\bar{r}_0^* = \sum r_{i,0}^*/k_0$ of the bootstrap sample $\{r_{i,0}^*\}$ obtained by re sampling R_0^m satisfies $Pr(|\bar{r}_0^* - \bar{r}_0| > \epsilon) \to 0$ for all $\epsilon > 0$.

Proof Define $g_n(r_{\cdot,0}) = r_{\cdot,0}$. From Lemma 3.3.1, $r_{\cdot,0}$ is asymptotically geometric with parameter P_{01}. This implies that $\mathrm{E}(g_n(r_{\cdot,0})) = 1/P_{01} < \infty$ and $\mathrm{Var}(g_n(r_{\cdot,0})) = (1 - P_{01})/P_{01}^2 < \infty$. Set $\zeta_n = \mathrm{E}(\bar{g}_n)$ and $\sigma_n^2 = \mathrm{Var}(\bar{g}_n)$, then by the CLT:

$$T_n = \frac{\bar{g}_n - \zeta_n}{\sigma_n} = \frac{\bar{r}_0 - 1/P_{01}}{\sqrt{(1 - P_{01})/nP_{01}^2}} = \frac{\sqrt{n}\left(\bar{r}_0 - \mu_{r_{\cdot,0}}\right)}{\sigma_{r_{\cdot,0}}} \xrightarrow{d} N(0, 1).$$

By Theorem 3.3.3, $G_n^*(\cdot)$ consistently estimates G_n. Hence, for any $\epsilon > 0$, $\Pr\left(|\bar{r}_0^* - \bar{r}_0| > \epsilon\right) \to 0$ ∎

The parameters P_{01} and P_{10} must be estimated to find values for the $G^2_{\max}$ statistic. The MLEs of these parameters are algebraically equivalent to the inverse of the bootstrap estimates $\bar{r}_0$ and $\bar{r}_1$. Define k_0 and k_1 to be the total number of runs of 0 and 1 necessary to achieve a bootstrap sequence of length n, then:

$$\hat{P}_{01} = \frac{n_{01}^n}{n_{01}^n + n_{00}^n} = \frac{k_0}{k_0 + \sum_{i=1}^{k_0}(r_{i,0} - 1)} = \frac{k_0}{\sum_{i=1}^{k_0} r_{i,0}} = \frac{1}{\bar{r}_0},$$
$$\hat{P}_{10} = \frac{n_{10}^n}{n_{10}^n + n_{11}^n} = \frac{k_1}{k_1 + \sum_{i=1}^{k_1}(r_{i,1} - 1)} = \frac{k_1}{\sum_{i=1}^{k_1} r_{i,1}} = \frac{1}{\bar{r}_1}.$$

The next lemma shows that the MLE of the mean of a geometric sequence is a consistent estimator. This lemma is specific to the set R_0^m, but can easily be modified for R_1^m.

Lemma 3.3.5 *The MLE $\hat{P}_{01} = 1/\bar{r}_0$ of the mean of an iid geometric random sample $\{r_{i,0} \mid i = 1, \ldots, k_0\}$ with parameter P_{01} is a consistent estimator of P_{01}.*

Proof By the central limit theorem:

$$\sqrt{n}(\bar{r}_0 - \mu) = \sqrt{n}(\bar{r}_0 - 1/P_{01}) \xrightarrow{d} \mathrm{N}(0, (1 - P_{01})P_{01}^{-2}) = \mathrm{N}(0, \sigma^2_{r_{\cdot,0}}).$$

Define $g(x) = 1/x$. Applying the delta method gives:

$$\sqrt{n}(1/\bar{r}_0 - P_{01}) = \sqrt{n}(g(\bar{r}) - g(1/P_{01})) \xrightarrow{d} \mathrm{N}(0, \sigma^2_{r_{\cdot,0}}(g'(1/P_{01}))^2) = \mathrm{N}(0, (1-P_{01})P_{01}^2).$$

Therefore, $\Pr(|\hat{P}_{01} - P_{01}| > \epsilon) = \Pr(|1/\bar{r} - P_{01}| > \epsilon) \to 0$ as $n \to \infty$. ∎

By Lemma 3.3.4, $\Pr\left(|\bar{r}^*_0 - \bar{r}_0| > 0\right) \to 0$, and by Lemma 3.3.5, $1/\bar{r}_0$ is a consistent estimator of P_{01}. Combining these facts gives the first main result of this section, which is summarized below.

Theorem 3.3.6 *The inverse of the bootstrap estimate $\bar{r}^*_0$ is a consistent estimator of P_{01}.*

Proof By Lemmas 3.3.4 and 3.3.5 we have:

$$\begin{aligned}
\Pr(|P^*_{01} - P_{01}| > \epsilon) &= \Pr(|P^*_{01} - \hat{P}_{01} + \hat{P}_{01} - P_{01}| > \epsilon) \\
&\leq \Pr(|P^*_{01} - \hat{P}_{01}| > \epsilon) + \Pr(|\hat{P}_{01} - P_{01}| > \epsilon) \\
&= \Pr(|1/\bar{r}^*_0 - 1/\bar{r}_0| > \epsilon) + \Pr(|\hat{P}_{01} - P_{01}| > \epsilon) \\
&= \Pr\left(\frac{|\bar{r}_0 - \bar{r}^*_0|}{\bar{r}_0\bar{r}^*_0} > \epsilon\right) + \Pr(|\hat{P}_{01} - P_{01}| > \epsilon) \\
&\to 0.
\end{aligned}$$

∎

The final justification of the bootstrap algorithm is to show that the resulting modified null likelihood function of the bootstrap algorithm is asymptotically equivalent to the modified null likelihood function of the original sequence. This will guarantee that the bootstrap values of $G^2_{\max}$ provide a good approximation of the true null distribution.

Theorem 3.3.7 *Using the bootstrap algorithm described in Section 3.3.1, under the hypothesis of no change, the modified likelihood function of the bootstrapped sequence y^* is asymptotically equivalent to the modified likelihood function of the original sequence y.*

Proof Suppose the number of runs in the original sequence is n_r and, without loss of generality, suppose y begins with an element of R_0^m, say $r_{0,0}$. The algorithm terminates when the length of the bootstrapped sequence of elements of $R^m = R_0^m \cup R_1^m$ is longer than the original sequence y. Let n_{r^*} denote the number of elements from R^m used in the bootstrapped sequence. Define f_{r^*} to be the distribution of the runs in R^m. The joint distribution of elements of R^m is:

$$f_{\vec{r}^*}(\vec{z}) = \begin{cases} \prod_{i=1}^{n_{r^*}/2} f_{r^*_{i,0}}(z_{i,0}) f_{r^*_{i,1}}(z_{i,1}) & \text{if } n_{r^*} \text{ is even,} \\ f_{r^*_{n_{r^*},0}}(z_{n_{r^*},0}) \prod_{i=1}^{(n_{r^*}-1)/2} f_{r^*_{i,0}}(z_{i,0}) f_{r^*_{i,1}}(z_{i,1}) & \text{if } n_{r^*} \text{ is odd.} \end{cases}$$

Let $f_p(z_0)$ denote the distribution of the initial value x_1 and define f_r to be the distribution of the runs of the original sequence y. The joint distribution of the runs of the original sequence y may be written as:

$$f_{\vec{r}}(\vec{z}) = \begin{cases} f_p(z_0) \prod_{i=1}^{n_r/2} f_{r_{i,0}}(z_{i,0}) f_{r_{i,1}}(z_{i,1}) & \text{if } n_r \text{ is even,} \\ f_p(z_0) f_{n_r,0}(z_{n_r,0}) \prod_{i=1}^{(n_r-1)/2} f_{r_{i,0}}(z_{i,0}) f_{r_{i,1}}(z_{i,1}) & \text{if } n_r \text{ is odd.} \end{cases}$$

Recall that the modified likelihood function L^* ignores the membership probability of the first term. Therefore, the $f_p(z_0)$ term in $f_y(\vec{z})$ is ignored in the modified likelihood function. By Lemmas 3.3.1 and 3.3.2 the asymptotic distributions of $f_{r_i,\cdot}$ and $f_{r_i^*,\cdot}$ are both geometric. From Theorems 3.2.5 and 3.3.6, the statistics $\hat{P}_{uv}$ and $\hat{P}^*_{uv}$ are both consistent for P_{uv}. Therefore, the asymptotic distribution of $f_{r_i^*,\cdot}$ is $f_{r_i,\cdot}$.

All that is left to show is that in the limit, the modified likelihood of $f_{\vec{r}^*}(\vec{y})$ is asymptotically equivalent to the modified likelihood of $f_r(y_0, \vec{y})$. The random selection of elements of R^m to construct the bootstrap sequence cause n_r and n_{r^*} to have the same order. Specifically, $\lim_{n\to\infty} \frac{n_r}{n_{r^*}} = 1$. Let $\tilde{f}_r(\vec{z})$ denote the joint *pdf* of the runs of y without the distribution

of the initial value x_1. Consider the ratio of the joint *pdf*s:

$$\lim_{n\to\infty}\left(\frac{f_{\vec{r}^*}(\vec{z})}{\tilde{f}_{\vec{r}}(\vec{z})}\right)=\left(\frac{\prod_{i=1}^{\infty} f_{r^*_{i,0}}(z_{i,0})f_{r^*_{i,1}}(z_{i,1})}{\prod_{i=1}^{\infty} f_{r_{i,0}}(z_{i,0})f_{r_{i,1}}(z_{i,1})}\right)=1.$$

Therefore, the modified likelihood of the bootstrapped sequence is asymptotically equivalent to the modified likelihood of the original sequence under the null hypothesis. ∎

Chapter 4

Simulations and Comparisons

Both the DCUSUM statistic and dependent LRT statistic provide detection and estimation methods for change points in dependent sequences of random variables. Asymptotically, both are valid choices, but the theoretical results do not indicate which method is preferred and under what circumstances. A variety of simulations are carried out to compare the DCUSUM test from Chapter 2 to the dependent LRT from Chapter 3. The performance of these tests is also compared to their independent counterparts discussed in Chapter 1 to illustrate the improvement of the generalized model for one step Markov dependence.

This chapter is organized as follows. In Section 4.2, the asymptotic results of the DCUSUM T_t statistic are reinforced and the performance is compared to the independent CUSUM test. Section 4.3 illustrates the asymptotic results of the dependent LRT G_t^2 statistic for fixed t and the performance of the statistic is compared to the independent LRT. In Section 4.4, the two proposed methods, DCUSUM and dependent LRT, are compared for size and power.

Four models were used for size comparisons and can be found in Table 4.1. The choice of parameters was used to simulate large, moderate, and small values of m, as well as one independent case. Since the population parameters for each model are known, the population values of m in Table 4.1 could be computed, using a tolerance of $tol = 0.01$. Specifically, m

was calculated as the value such that:

$$\left| \boldsymbol{P}^m - \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\pi} \end{pmatrix} \right| < tol = 0.01,$$

where $|\cdot|$ denotes component-wise absolute differences. The values of $l = 1/20$ and $h = 1 - 1/20$ were chosen to use the largest justifiable amount of the sequence. More details about the choices of l and h can be found in Miller and Siegmund [26].

Five models were used for power comparisons and can be found in Table 4.2. These models were selected to demonstrate large, moderate, and small changes in the parameters $p(1), p(2), P_{11}(1)$ and $P_{11}(2)$.

Before the results are presented, it is important to note that two issues arose in some of the small sample simulations. First, it was possible that the value of $\hat{m}$ was larger than the permissible lower bound for $t_1 = nl$ and/or the lower bound for $n - t_2 = nh$. In order for the variance to be computed correctly, $\hat{m}$ was redefined in all simulations to be the minimum of the set $\{\hat{m}, nl, n(1-h)\}$. When $\hat{m}$ was redefined, an overestimation of the variance of the statistics occurred. Second, the large values of P_{11} may have caused some sequences to contain all zeros, all ones, or too few changes to compute the values of the statistics. This caused situations where the MLEs were not able to be estimated, due to zeros in the denominator.

For all results, 2000 iterations were run, and the cases where the statistics could not be computed were removed. The 2000 iterations should provide reasonable accuracy to two decimal places and are only used to compare the methods, gain insight on method preference, and provide an alternative verification of theoretical results.

Table 4.1: Parameters for Null Simulations

p	P_{11}	m	Description
0.4	0.9	22	Large (L)
0.7	0.9	10	Moderate (M)
0.7	0.75	2	Small (S)
0.7	0.7	0	Independent (I)

Table 4.2: Parameters for Power Comparisons

$p(1)$	$P_{11}(1)$	$p(2)$	$P_{11}(2)$	τ	Description
0.8	0.9	0.2	0.9	$(2/5)n$	Large Change in p (L_p)
0.8	0.9	0.4	0.9	$(2/5)n$	Moderate Change in p (M_p)
0.8	0.9	0.6	0.9	$(2/5)n$	Small Change in p (S_p)
0.8	0.9	0.4	0.5	$(2/5)n$	Moderate Change in p and P_{11} ($\mathrm{M}_{p,P_{11}}$)
0.8	0.9	0.6	0.7	$(2/5)n$	Small Change in p and P_{11} ($\mathrm{S}_{p,P_{11}}$)

4.1 Estimating m with Unknown Parameters

In practice, the population parameters p and P_{11} are unknown and must be estimated. Subsequently, the value of m must also be estimated from the data. This estimation is discussed below.

Under the null model of no change, the values of p and P_{11} are estimated using the modified MLEs $\hat{p}$ and $\hat{P}_{11}$ defined in equation (3.5). The value of m for a tolerance of 0.01 is then estimated using the modified MLEs as follows:

$$\hat{m} = \min\left\{ t : \left| \hat{\boldsymbol{P}}^t - \begin{pmatrix} \hat{\boldsymbol{\pi}} \\ \\ \hat{\boldsymbol{\pi}} \end{pmatrix} \right| < tol = 0.01 \right\},$$

where $\hat{\boldsymbol{P}}^t$ is the transition matrix comprised of the modified MLEs with a change point at t, and $\hat{\boldsymbol{\pi}} = (1-\hat{p}\ \ \hat{p})$ is the modified MLE of the stationary distribution.

In the presence of small samples and large values of $\hat{m}$, the assumption that $\hat{m} < \min\{t_1, n - t_2\}$ for $nl \leq t_1 < t_2 \leq nh$ may be violated. If this occurs, the value of $\hat{m}$

is taken to be $\hat{m}' = \min\{t_1, n - t_2\}$ to ensure that the variance calculations of the test statistics are still valid. When $\hat{P}_{11} > \hat{p}$, this causes an overestimation of the variance of the test statistic.

4.2 DCUSUM Simulations

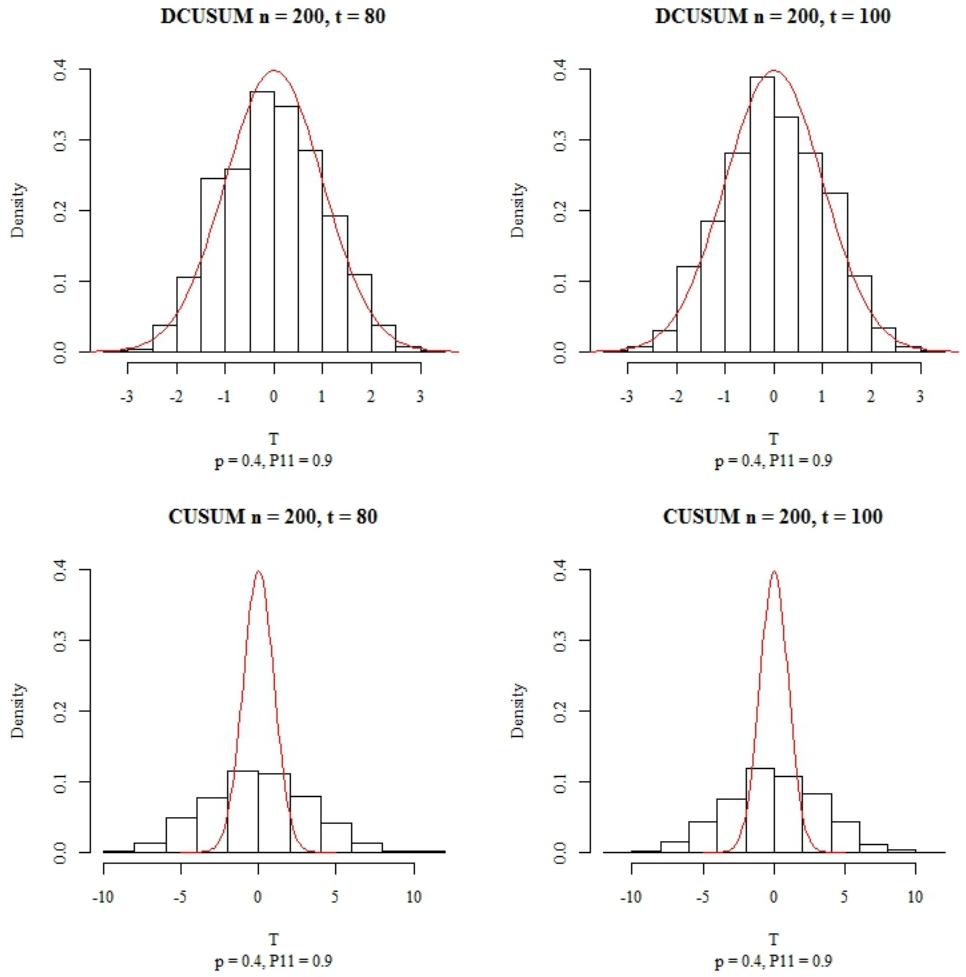
The generalization of the CUSUM statistic to include the one step Markov dependence assumption resulted in the DCUSUM statistic defined by equation (2.2). The test statistic $T^2_{\max}$ defined by equation (2.7) was used to determine the existence and location of a change point τ. The asymptotic distribution of T_t and the size of the DCUSUM test are discussed below.

4.2.1 Sampling Distribution of T_t for Fixed t

In the proof of Lemma 2.4.1, the asymptotic distribution of T_t for fixed t was found to be $\mathrm{N}(0, 1)$. The sampling distributions of T_{80} and T_{100} for $n = 200$ are shown in Figures 4.1, 4.2, and 4.3 with 2000 simulated values for each. These plots both reinforce the standard normal distribution of the T_t statistic for fixed t when the DCUSUM statistic was used, as well as illustrate the inflated variance of the T_t statistic resulting from the CUSUM statistic when $\hat{m}$ was large.

Table 4.3 contains the means and standard deviations of the sampling distributions for both DCUSUM and CUSUM statistics. Under the assumption that $P_{11} > p$, the variance of the CUSUM statistic is less than the variance of the DCUSUM statistic. In the presence of m-dependence, this caused an underestimation of the variance of the CUSUM statistic when the dependence is ignored. An underestimation of the variance of the CUSUM statistic lead to an overestimation of the variance of the T_t statistic. Table 4.3 reinforces this claim as the standard deviations of the T_t statistics when using CUSUM are larger than 1, even in the case when $\hat{m}$ is small (S). When the variables in the sequence are independent, the mean

Figure 4.1: Sampling Distribution of T_{80} and T_{100}, Simulation (L)



and standard deviation of the T_t statistics resulting from DCUSUM and CUSUM are equal.

Figure 4.2: Sampling Distribution of T_{80} and T_{100}, Simulation (M)

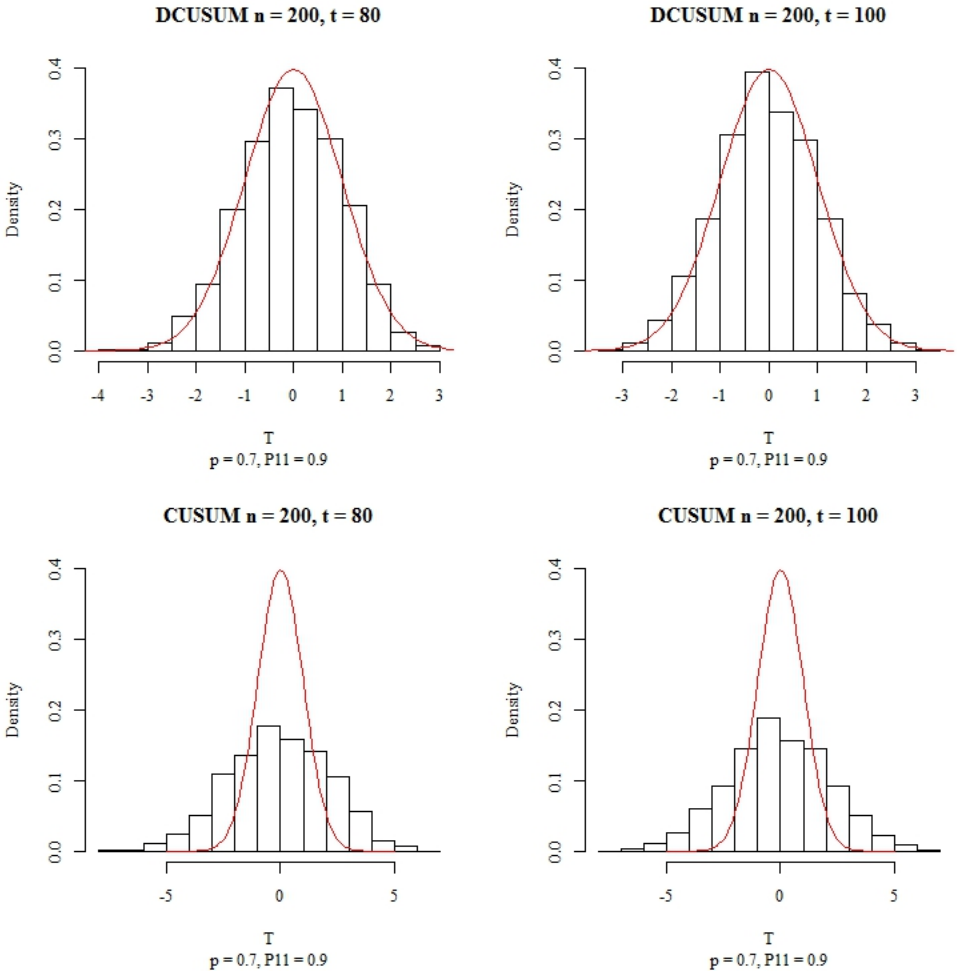
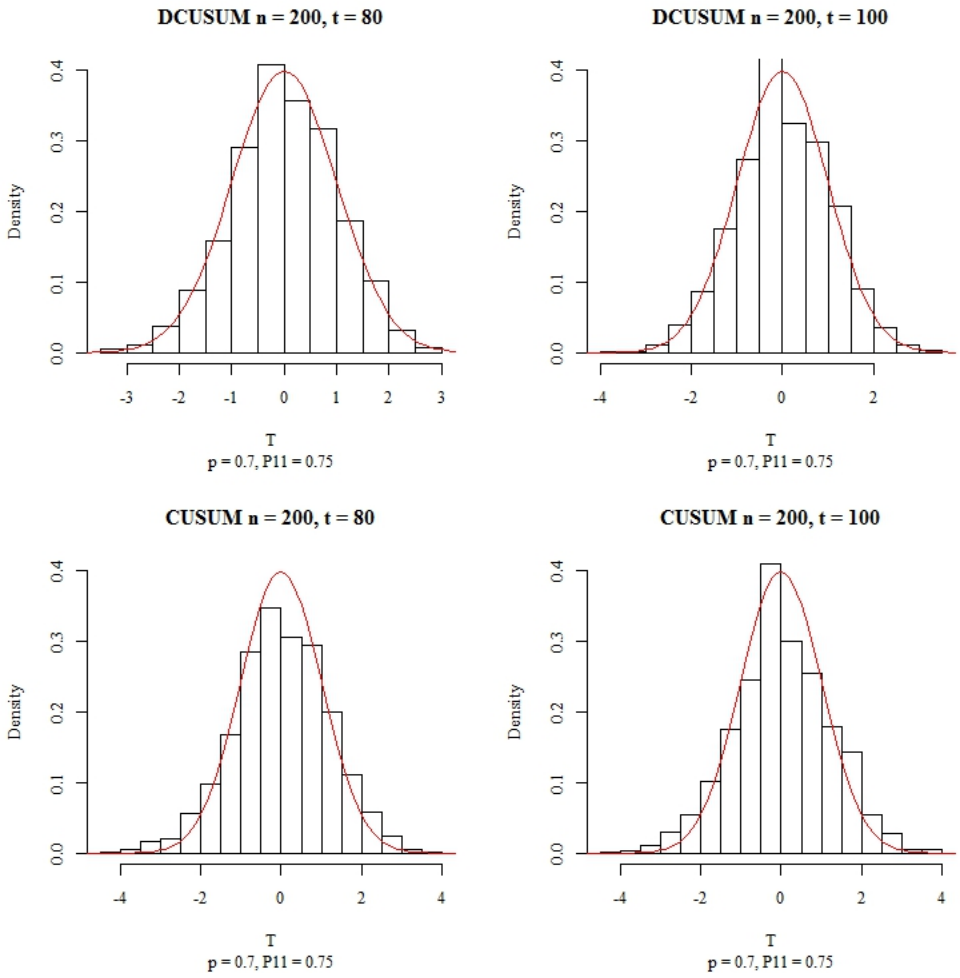


Table 4.3: Sample Mean and Standard Deviation of T_t for $n = 200$

		DCUSUM		CUSUM	
t	Model	mean	st. dev	mean	st. dev
80	L	-0.014	1.036	-0.053	3.255
	M	-0.033	1.018	-0.067	2.208
	S	0.012	1.002	0.015	1.187
	I	-0.028	0.993	-0.026	0.994
100	L	0.011	1.037	0.026	3.257
	M	-0.036	1.020	-0.073	2.210
	S	0.031	1.017	0.034	1.206
	I	-0.035	0.994	-0.033	0.996

Figure 4.3: Sampling Distribution of T_{80} and T_{100}, Simulation (S)



4.2.2 Size Comparison DCUSUM

The next simulation compares the sizes of the DCUSUM and CUSUM tests under H_0 using the approximation of the tail probabilities given in equation (1.14). The size was determined by generating 2000 sequences and counting the number of p-values or upper bounds that were less than or equal to 0.05.

The simulated size comparison may be found in Table 4.4. As expected, the sizes of the DCUSUM and CUSUM tests are similar in the independent case. As $\hat{m}$ grows, the size of the CUSUM test grows to an unreasonable level. This is due to the underestimation of the variance of the CUSUM statistic in the presence of m-dependence.

For small samples ($n \leq 150$) with small to moderate $\hat{m}$ values, the Worsley upper bound method provides sizes close to 0.05, while the Brownian approximation method is more conservative with almost all sizes less than 0.05. As the sample size grows ($n \geq 500$), the size of the Worsley upper bound method changes roles with the size of the Brownian approximation method. When the value of $\hat{m}$ is large, both tests display comparable sizes.

As expected, the large sample performance of the Brownian approximation method is acceptable, with the exception of the large difference case (L). As n grows, both the Worsley upper bound method and Brownian approximation method provide conservative tests. While the size of the Brownian approximation method does not approach 0.05, it provides a conservative test that is more liberal than the Worsley upper bound method.

From these results, particularly models (M) and (S), the Brownian approximation method is recommended for larger samples ($n \geq 500$), while the Worsley upper bound method is recommended for smaller samples ($n < 500$). When $\hat{m}$ is large, the two tests have similar size. In this case, the Brownian approximation method is recommended due to faster run time than the Worsley upper bound method. In any case, the sizes of both DCUSUM tests substantially outperform the size of the CUSUM test in the presence of m-dependence.

Table 4.4: Size Comparison of DCUSUM (Depdent) and CUSUM (Independent) Procedures

n		Brownian DCUSUM	Worsley DCUSUM	Brownian CUSUM
50	L	0.04	0.11	0.83
	M	0.05	0.08	0.60
	S	0.02	0.03	0.08
	I	0.02	0.03	0.03
75	L	0.03	0.06	0.90
	M	0.03	0.05	0.65
	S	0.02	0.03	0.09
	I	0.02	0.02	0.02
100	L	0.02	0.05	0.93
	M	0.03	0.05	0.70
	S	0.03	0.03	0.10
	I	0.03	0.03	0.03
150	L	0.01	0.03	0.97
	M	0.03	0.04	0.75
	S	0.03	0.03	0.12
	I	0.02	0.02	0.03
200	L	0.01	0.02	0.98
	M	0.03	0.04	0.79
	S	0.04	0.03	0.12
	I	0.02	0.02	0.03
250	L	0.01	0.03	0.99
	M	0.03	0.04	0.82
	S	0.04	0.03	0.13
	I	0.03	0.02	0.04
300	L	0.01	0.03	0.98
	M	0.03	0.04	0.84
	S	0.03	0.02	0.12
	I	0.03	0.02	0.04
500	L	0.01	0.02	0.99
	M	0.03	0.03	0.86
	S	0.03	0.02	0.13
	I	0.04	0.02	0.04
1000	L	0.03	0.03	0.99
	M	0.04	0.03	0.88
	S	0.03	0.01	0.14
	I	0.05	0.02	0.05

4.2.3 Power Comparison DCUSUM

The goal of this section is to determine which of the two methods, Brownian approximation to the p-value or the Worsley upper bound, has higher power and under what conditions. Five models were simulated under the alternative hypothesis of one change, and are described in Table 4.2. Three change point locations of $\tau = \frac{n}{5}, \frac{2n}{5},$ and $\frac{n}{2}$ were used. The construction of the $T^2_{\max}$ statistic is symmetric, so there was no need to use change points after the midpoint of the data set.

The approximate power calculations (APCs) were found by simulating 2000 sequences in each setting and finding the proportion of approximate p-values or upper bounds that were less than or equal to 0.05. The APCs are a measure of the empirical power of each method, and may be found in Tables 4.5, 4.6, 4.7.

For small samples ($n \leq 100$), the results are inconsistent and difficult to interpret. This is due to the fact that the value of $\hat{m}$ may be larger than some permissible values of t_1 or $n - t_2$, or that the sequences may contain too few runs for the $T^2_{\max}$ statistic to be computed. Because of this, only the results with $n \geq 150$ are discussed.

It is well known that the CUSUM test has higher power for change point locations near the middle of the sequence and lower power for locations at the tails. This fact appears to be true for DCUSUM methods as well, and is reinforced by the APC values. For $n \geq 150$, the APC for each change point location (APC_τ) and each method satisfies $\text{APC}_{\frac{n}{5}} \leq \text{APC}_{\frac{2n}{5}} \leq \text{APC}_{\frac{n}{2}}$.

The APC values tell a similar story as the size simulations. Regardless of the alternative model, when the sample size is reasonably small ($n \leq 200$), the Worsley upper bound method provides a more powerful test than the Brownian approximation method. For moderate sample sizes ($250 \leq n \leq 500$) the Worsley upper bound method is more powerful when only the parameter p changes, while the Brownian approximation method is more powerful when both parameters p and P_{11} change. When large samples are available ($n \geq 1000$), the Brownian approximation method is equivalent to or more powerful than the Worsley upper bound method for all simulated models.

The recommended models are the same as in Section 4.2.2. For smaller samples ($n \leq 200$), the Worsley upper bound method should be used over the Brownian approximation method. When moderate samples are available ($250 \leq n \leq 500$), either method may be used, unless there is some prior knowledge about the possible alternative model. For large samples ($n \geq 1000$), the Brownian approximation method is preferred due to higher power and shorter run time.

Table 4.5: APC of DCUSUM Tests, $\tau = (1/5)n$

n	Model	Brownian DCUSUM	Worsley DCUSUM
50	L_p	0.23	0.33
	M_p	0.07	0.11
	S_p	0.02	0.04
	$M_{p,P_{11}}$	0.07	0.13
	$S_{p,P_{11}}$	0.01	0.02
75	L_p	0.23	0.35
	M_p	0.05	0.10
	S_p	0.02	0.04
	$M_{p,P_{11}}$	0.20	0.28
	$S_{p,P_{11}}$	0.03	0.04
100	L_p	0.25	0.36
	M_p	0.03	0.08
	S_p	0.02	0.04
	$M_{p,P_{11}}$	0.38	0.43
	$S_{p,P_{11}}$	0.05	0.06
150	L_p	0.27	0.40
	M_p	0.05	0.10
	S_p	0.02	0.03
	$M_{p,P_{11}}$	0.62	0.63
	$S_{p,P_{11}}$	0.10	0.11
200	L_p	0.33	0.45
	M_p	0.06	0.11
	S_p	0.02	0.04
	$M_{p,P_{11}}$	0.77	0.76
	$S_{p,P_{11}}$	0.18	0.17
250	L_p	0.38	0.49
	M_p	0.08	0.13
	S_p	0.03	0.04
	$M_{p,P_{11}}$	0.86	0.85
	$S_{p,P_{11}}$	0.26	0.24
300	L_p	0.46	0.57
	M_p	0.12	0.18
	S_p	0.04	0.05
	$M_{p,P_{11}}$	0.93	0.92
	$S_{p,P_{11}}$	0.32	0.28
500	L_p	0.77	0.82
	M_p	0.27	0.31
	S_p	0.07	0.08
	$M_{p,P_{11}}$	0.99	0.99
	$S_{p,P_{11}}$	0.56	0.48
1000	L_p	1.00	1.00
	M_p	0.80	0.80
	S_p	0.22	0.18
	$M_{p,P_{11}}$	1.00	1.00
	$S_{p,P_{11}}$	0.89	0.83

Table 4.6: APC of DCUSUM Tests, $\tau = (2/5)n$

n	Model	Brownian DCUSUM	Worsley DCUSUM
50	L_p	0.03	0.21
	M_p	0.03	0.08
	S_p	0.03	0.06
	$M_{p,P_{11}}$	0.10	0.19
	$S_{p,P_{11}}$	0.02	0.06
75	L_p	0.06	0.28
	M_p	0.02	0.10
	S_p	0.02	0.05
	$M_{p,P_{11}}$	0.31	0.40
	$S_{p,P_{11}}$	0.06	0.10
100	L_p	0.13	0.36
	M_p	0.04	0.14
	S_p	0.04	0.07
	$M_{p,P_{11}}$	0.47	0.54
	$S_{p,P_{11}}$	0.11	0.14
150	L_p	0.32	0.52
	M_p	0.11	0.20
	S_p	0.05	0.08
	$M_{p,P_{11}}$	0.75	0.77
	$S_{p,P_{11}}$	0.19	0.21
200	L_p	0.51	0.65
	M_p	0.21	0.30
	S_p	0.09	0.11
	$M_{p,P_{11}}$	0.88	0.88
	$S_{p,P_{11}}$	0.30	0.30
250	L_p	0.64	0.74
	M_p	0.29	0.38
	S_p	0.10	0.12
	$M_{p,P_{11}}$	0.95	0.95
	$S_{p,P_{11}}$	0.37	0.36
300	L_p	0.77	0.83
	M_p	0.40	0.48
	S_p	0.14	0.17
	$M_{p,P_{11}}$	0.98	0.98
	$S_{p,P_{11}}$	0.46	0.43
500	L_p	0.98	0.98
	M_p	0.76	0.78
	S_p	0.28	0.28
	$M_{p,P_{11}}$	1.00	1.00
	$S_{p,P_{11}}$	0.73	0.68
1000	L_p	1.00	1.00
	M_p	0.99	0.99
	S_p	0.61	0.56
	$M_{p,P_{11}}$	1.00	1.00
	$S_{p,P_{11}}$	0.97	0.94

Table 4.7: APC of DCUSUM Tests, $\tau = (1/2)n$

n	Model	Brownian DCUSUM	Worsley DCUSUM
50	L_p	0.02	0.08
	M_p	0.02	0.07
	S_p	0.027	0.05
	$M_{p,P_{11}}$	0.03	0.08
	$S_{p,P_{11}}$	0.08	0.18
75	L_p	0.04	0.21
	M_p	0.03	0.11
	S_p	0.03	0.08
	$M_{p,P_{11}}$	0.07	0.12
	$S_{p,P_{11}}$	0.28	0.38
100	L_p	0.13	0.38
	M_p	0.07	0.19
	S_p	0.05	0.09
	$M_{p,P_{11}}$	0.13	0.16
	$S_{p,P_{11}}$	0.48	0.55
150	L_p	0.41	0.58
	M_p	0.19	0.29
	S_p	0.09	0.13
	$M_{p,P_{11}}$	0.22	0.23
	$S_{p,P_{11}}$	0.75	0.78
200	L_p	0.62	0.74
	M_p	0.31	0.41
	S_p	0.12	0.16
	$M_{p,P_{11}}$	0.29	0.29
	$S_{p,P_{11}}$	0.88	0.88
250	L_p	0.77	0.83
	M_p	0.43	0.50
	S_p	0.17	0.20
	$M_{p,P_{11}}$	0.39	0.38
	$S_{p,P_{11}}$	0.96	0.96
300	L_p	0.85	0.89
	M_p	0.54	0.60
	S_p	0.21	0.23
	$M_{p,P_{11}}$	0.48	0.45
	$S_{p,P_{11}}$	0.98	0.98
500	L_p	0.99	0.99
	M_p	0.84	0.84
	S_p	0.38	0.38
	$M_{p,P_{11}}$	0.73	0.68
	$S_{p,P_{11}}$	1.00	1.00
1000	L_p	1.00	1.00
	M_p	0.99	0.99
	S_p	0.71	0.66
	$M_{p,P_{11}}$	0.98	0.96
	$S_{p,P_{11}}$	1.00	1.00

4.3 Dependent LRT Simulations

The generalization of the LRT defined in Section 1.2.3 for an m-dependent sequence is the dependent LRT, as discussed in Chapter 3. Large values of the log likelihood ratio statistic $G^2_{\max}$ defined by equation (3.6) provide evidence of the existence of a change point τ.

4.3.1 Sampling Distribution of G_t^2 for Fixed t

In Section 3.2 the asymptotic distribution of G_t^2 was shown to be χ^2_2 for any fixed value of t. This result is reinforced by simulations from this section.

The sampling distributions of each model in Table 4.1 for fixed times $t = 80$ and 100 for $n = 200$ are shown in Figures 4.4, 4.5, 4.6, and 4.7 with 2000 simulated values for each. The red curve superimposed on Figures 4.4, 4.5, and 4.6 is the density of a χ^2_2 random variable while the curve on Figure 4.7 is the χ^2_1 density. In the three cases where m-dependence is present (L, M, S), the distribution of G_t^2 under the dependent assumption appear to follow a χ^2_2 distribution, while the distribution of G_t^2 under the independent assumption have much heavier tails.

As $\hat{m}$ grows, the asymptotic distribution of the independent G_t^2 statistic deviates further from the χ^2_2 distribution. This can be seen clearly by the sample percentiles in Tables 4.8, 4.9, and 4.10. This indicates that an overestimation of the variance of the independent G_t^2 statistic occurs when the variables in the sequence display m-dependence, while the dependent G_t^2 statistic is not severely affected by the increase in the value of $\hat{m}$.

In simulation (I), the true distribution is χ^2_1. Due to the unnecessary estimation of the nuisance parameter P_{11}, the dependent LRT still behaves like a χ^2_2 random variable. Clearly, from Table 4.11, the independent LRT is much closer to the true asymptotic distribution than the dependent LRT when m-dependence is not present.

Figure 4.4: Sampling Distribution of G^2_{80} and G^2_{100}, Simulation (L)

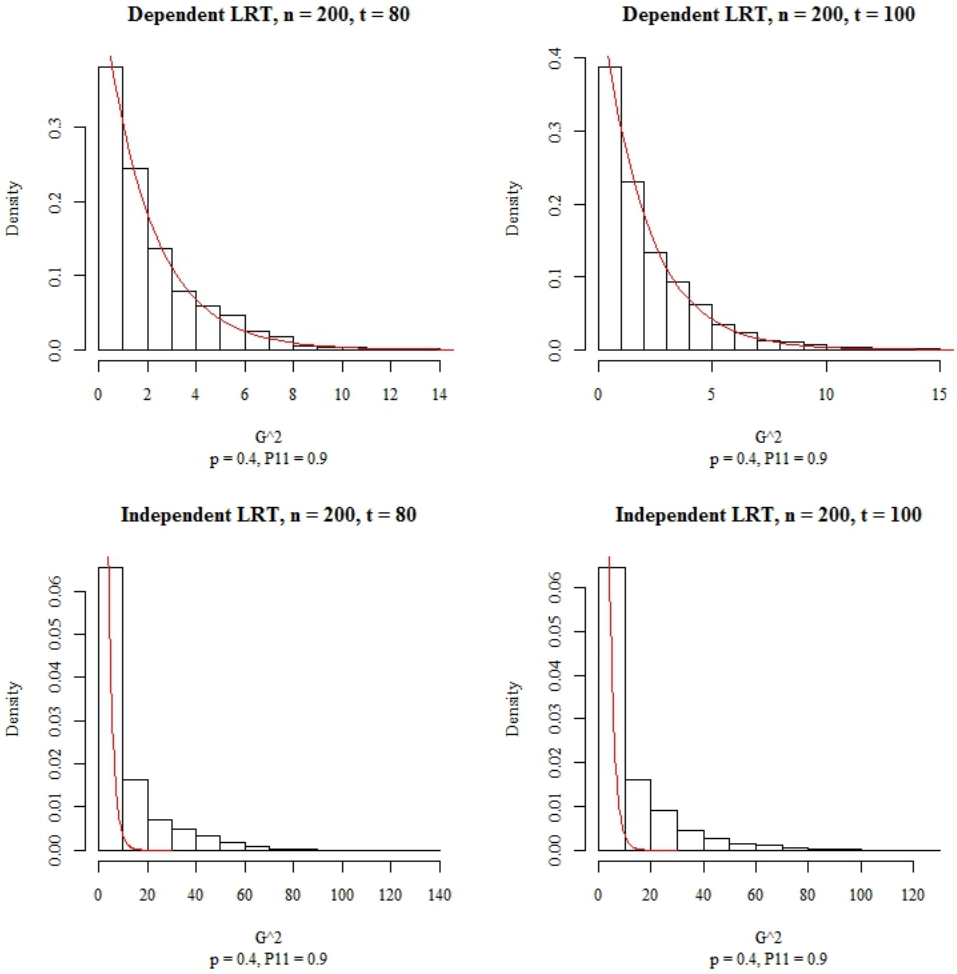


Table 4.8: Sample Percentiles of G^2_t for $n = 200$ (L)

t	Percentile	Actual χ^2_2	Dependent LRT	Independent LRT
80	P_{90}	4.6052	5.0723	33.5870
	P_{95}	5.9915	6.1928	43.2269
	P_{99}	9.2103	8.9067	66.3639
100	P_{90}	4.6052	4.9627	30.8672
	P_{95}	5.9915	6.4071	43.7261
	P_{99}	9.2103	9.8651	69.3053

Figure 4.5: Sampling Distribution of G^2_{80} and G^2_{100}, Simulation (M)

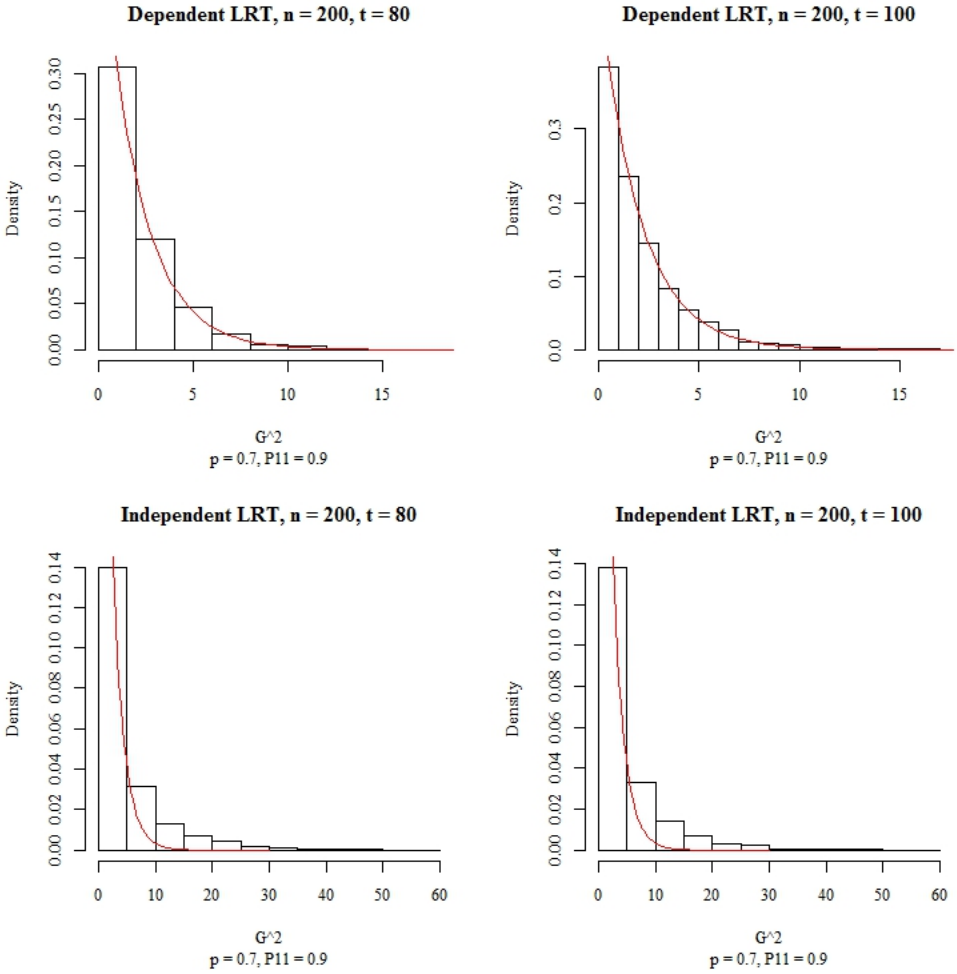


Table 4.9: Sample Percentiles of G^2_t for $n = 200$ (M)

t	Percentile	Actual χ^2_2	Dependent LRT	Independent LRT
80	P_{90}	4.6052	4.7322	12.4881
	P_{95}	5.9915	6.2883	18.8869
	P_{99}	9.2103	9.9770	30.1982
100	P_{90}	4.6052	5.0069	12.3393
	P_{95}	5.9915	6.4603	18.1146
	P_{99}	9.2103	9.8181	30.1177

Figure 4.6: Sampling Distribution of G^2_{80} and G^2_{100}, Simulation (S)

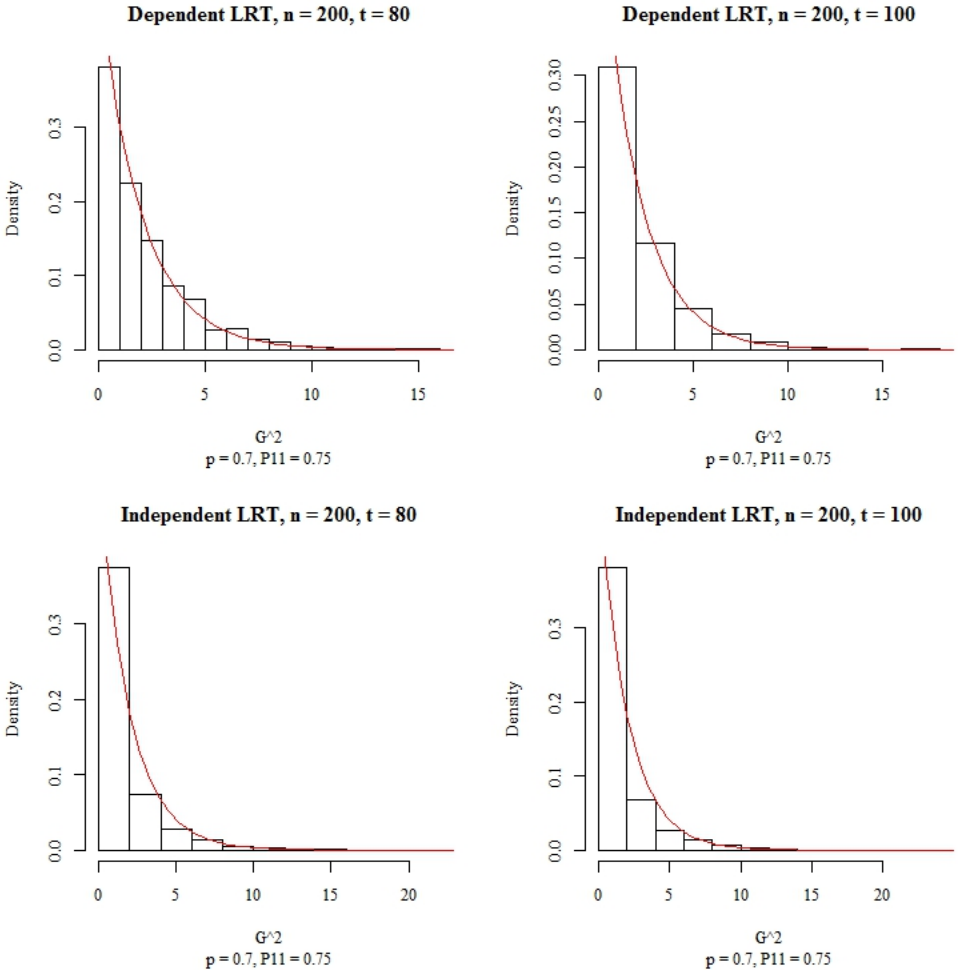


Table 4.10: Sample Percentiles of G^2_t for $n = 200$ (S)

t	Percentile	Actual χ^2_2	Dependent LRT	Independent LRT
80	P_{90}	4.6052	4.8394	4.0587
	P_{95}	5.9915	6.5930	5.8008
	P_{99}	9.2103	9.4637	11.0458
100	P_{90}	4.6052	4.9335	4.0095
	P_{95}	5.9915	6.3966	5.8418
	P_{99}	9.2103	9.9148	9.7403

Figure 4.7: Sampling Distribution of G^2_{80} and G^2_{100}, Simulation (I)

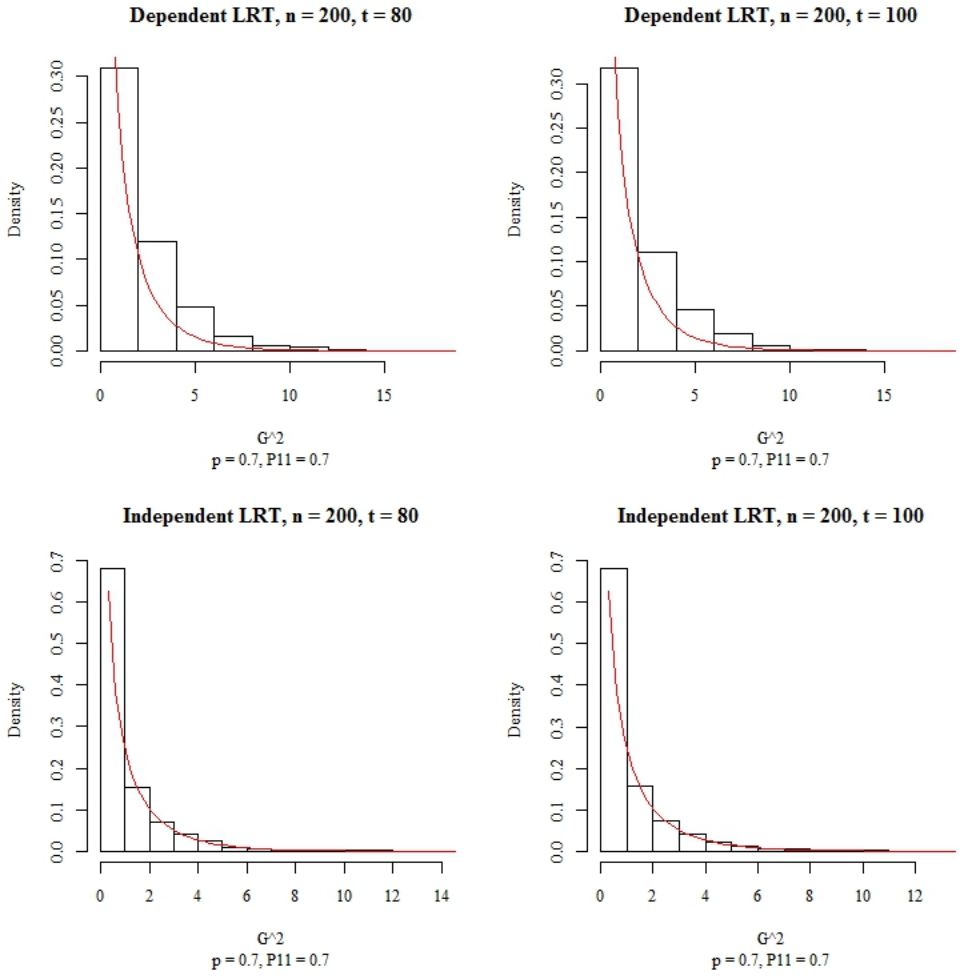


Table 4.11: Sample Percentiles of G^2_t for $n = 200$ (I)

t	Percentile	Actual χ^2_1 [a]	Dependent LRT	Independent LRT
80	P_{90}	2.7055	4.7583	2.9204
	P_{95}	3.8415	5.9989	4.2642
	P_{99}	6.6349	9.8130	7.9456
100	P_{90}	2.7055	4.8756	2.8659
	P_{95}	3.8415	6.1566	4.0134
	P_{99}	6.6349	8.6099	6.8867

[a] Under the independent assumption, $G^2_t \sim \chi^2_1$

Table 4.12: Sample Percentiles of $G^2_{\max}$

n	Percentile	χ^2_6	χ^2_7	(L)	(M)	(S)
500	P_{90}	10.6446	12.0170	11.4262	10.8239	10.9749
	P_{95}	12.5916	14.0671	13.2418	12.3801	12.5361
	P_{99}	16.8119	18.4753	16.5208	15.8935	15.5650
1000	P_{90}	10.6446	12.0170	11.4574	11.5709	11.5989
	P_{95}	12.5916	14.0671	13.0314	13.6001	13.1544
	P_{99}	16.8119	18.4753	16.8596	17.4270	16.8833
2000	P_{90}	10.6446	12.0170	11.9689	11.9156	11.6701
	P_{95}	12.5916	14.0671	13.9258	13.5675	13.6609
	P_{99}	16.8119	18.4753	17.7128	17.2436	18.0996

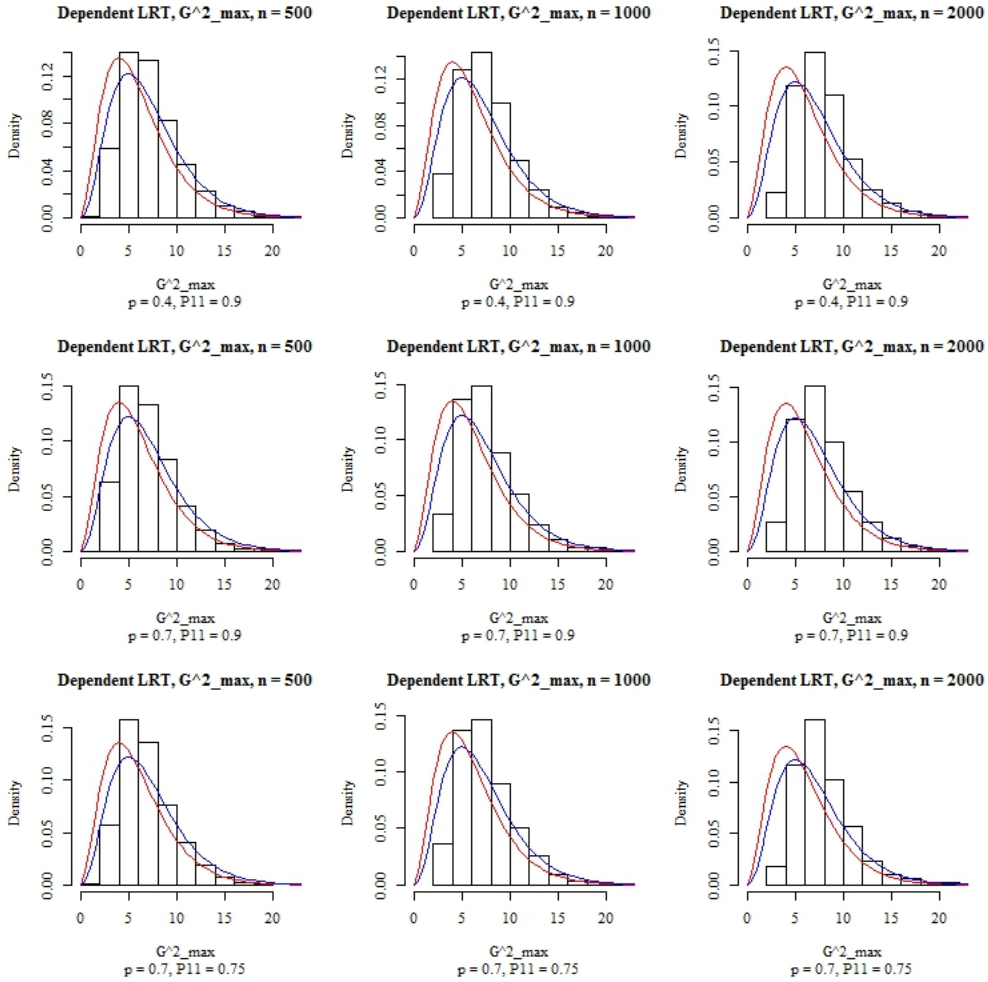
4.3.2 Approximate Asymptotic Distribution of $G^2_{\max}$

In Section 3.2, the asymptotic distribution of the dependent likelihood ratio statistic G^2_t for fixed t was shown to follow a χ^2_2 distribution. The added complexity of the m-dependence assumption for the sequence y caused difficulty in obtaining the asymptotic distribution of the $G^2_{\max}$ statistic. Instead, simulated tail distributions are used to approximate the asymptotic distribution. The results here are similar to the results found in Hinkley [12] and Feder [9], except that the asymptotic tail distribution for m-dependent sequences was found to be bounded (approximately) below by a χ^2_6 and above by a χ^2_7 random variable.

The models (L), (M), and (S) were simulated with $n = 500, 1000$, and 2000, and the $90^{th}, 95^{th}$, and 99^{th} percentiles of $G^2_{\max}$ for 2000 repetitions are displayed in Table 4.12. While it seems that the percentiles of $G^2_{\max}$ are effected by the choice of parameters p and P_{11}, the percentiles do not fall outside of the bounds of the χ^2_6 and χ^2_7 percentiles for any of the values of n.

Histograms of the distributions are shown in Figure 4.8. The red curve is the density of a χ^2_6 random variable, and the blue curve is the density of a χ^2_7 random variable. The density of simulated values near the middle of the data is larger than the χ^2_7 distribution, but the tail values fall between the two bounds.

Figure 4.8: Sampling Distribution of $G^2_{\max}$



4.3.3 Size Comparison and Bootstrap Effectiveness for LRT

The next simulation compares the sizes of the independent and dependent LRT under H_0 using the approximation of the tail probabilities given in equation (1.14) and the bootstrap method described in Section 3.3.1. The parameter values for each simulated model are defined in Table 4.1. A description of how the size was computed can be found in the first paragraph of Section 4.2.2.

As mentioned in Chapter 3, the bootstrap method requires a reasonably large sample size to provide approximate p-values. For each simulation in Table 4.13, the sizes do indeed approach 0.05, but at very different rates. The sample size necessary for Case (S) with small $\hat{m}$ to display sizes close to 0.05 is much smaller than the sample sizes necessary for Cases (M) and (L) with larger $\hat{m}$. Overall, the dependent LRT provides a liberal test, with a majority of sizes larger than 0.05.

The size of the bootstrap method in all three m-dependent cases outperforms the naive approach of the independent statistic when m-dependence is present. This is shown in Table 4.13.

Asymptotically, the bootstrap algorithm for the dependent LRT was justified in Section 3.3.3. Figures 4.9 and 4.10 provide a different form of justification of the bootstrap algorithm through simulation.

When the null hypothesis is true, the p-values generated via a valid procedure should follow a uniform distribution on the interval $(0, 1)$. The small value of $\hat{m}$ using model (S) displayed in Figure 4.9 shows that the bootstrap p-values approach a uniform distribution for samples as small as $n = 250$. The larger value of $\hat{m}$ in Figure 4.10 also shows the convergence to a uniform distribution, but illustrates that a larger sample size is necessary. Both Figures 4.9 and 4.10 show that the p-values generated from the independent LRT are inappropriate regardless of the size of $\hat{m}$.

For small samples ($n \leq 75$) and the large value of $\hat{m}$ in model (L), the size results are inconsistent. With large samples ($n \geq 100$), the sizes improve as expected. The larger the

Table 4.13: Size Comparison of Bootstrap and Independent LRT

n		Bootstrap LRT	Independent LRT
50	L	0.12	0.74
	M	0.23	0.50
	S	0.13	0.07
	I	0.05	0.02
75	L	0.22	0.86
	M	0.23	0.58
	S	0.11	0.08
	I	0.05	0.03
100	L	0.25	0.92
	M	0.19	0.68
	S	0.08	0.10
	I	0.05	0.03
150	L	0.24	0.97
	M	0.13	0.77
	S	0.07	0.12
	I	0.04	0.04
200	L	0.17	0.99
	M	0.11	0.80
	S	0.06	0.14
	I	0.05	0.04
250	L	0.14	0.99
	M	0.09	0.84
	S	0.07	0.13
	I	0.05	0.05

value of $\hat{m}$ the larger the sample size necessary to achieve a size of 0.05. For all cases where dependence is present and the sample size is reasonable ($n \geq 100$), the huge sizes of the independent statistic reinforce the fact that this statistic is not appropriate.

4.4 DCUSUM and Dependent LRT Comparison

This section contains simulations to determine the conditions when the DCUSUM p-value approximation, DCUSUM Worsley upper bound, or dependent LRT are preferable. The size and power of the DCUSUM methods and dependent LRT procedures are compared using the models is Tables 4.1 and 4.2.

4.4.1 Size Comparison

The sizes of the DCUSUM Brownian approximation to the p-value, DCUSUM Worsley upper bound, and dependent LRT bootstrap p-values are given in Tables 4.4 and 4.13. For the models (S), (M), and (L), where m-dependence is present, the size of the dependent LRT is much larger than both the DCUSUM Brownian approximation and the DCUSUM Worsley upper bound. As the sample size grows, the discrepancy in size decreases.

If size is a priority, the recommended method is dependent on sample size. For small to moderate samples ($n \leq 200$), one of the DCUSUM procedures is recommended. If the sample size is large, ($n \geq 250$), then the methods provide similar sizes. In this case, the DCUSUM procedures are conservative, while the dependent LRT procedure is liberal.

4.4.2 Power Comparison

The powers of the DCUSUM Brownian approximation method, DCUSUM Worsley upper bound method, and dependent LRT are compared in this section. This is done in two ways. One comparison uses the approximate power calculations (APCs) of the two models (M_p) and ($\mathrm{S}_{p,P_{11}}$) to compare the power of each method in practice. The other comparison calculates

Figure 4.9: Histograms of p-values for G^2 Bootstrap and Independent p-values for Simulation (S) with Varying Sample Sizes

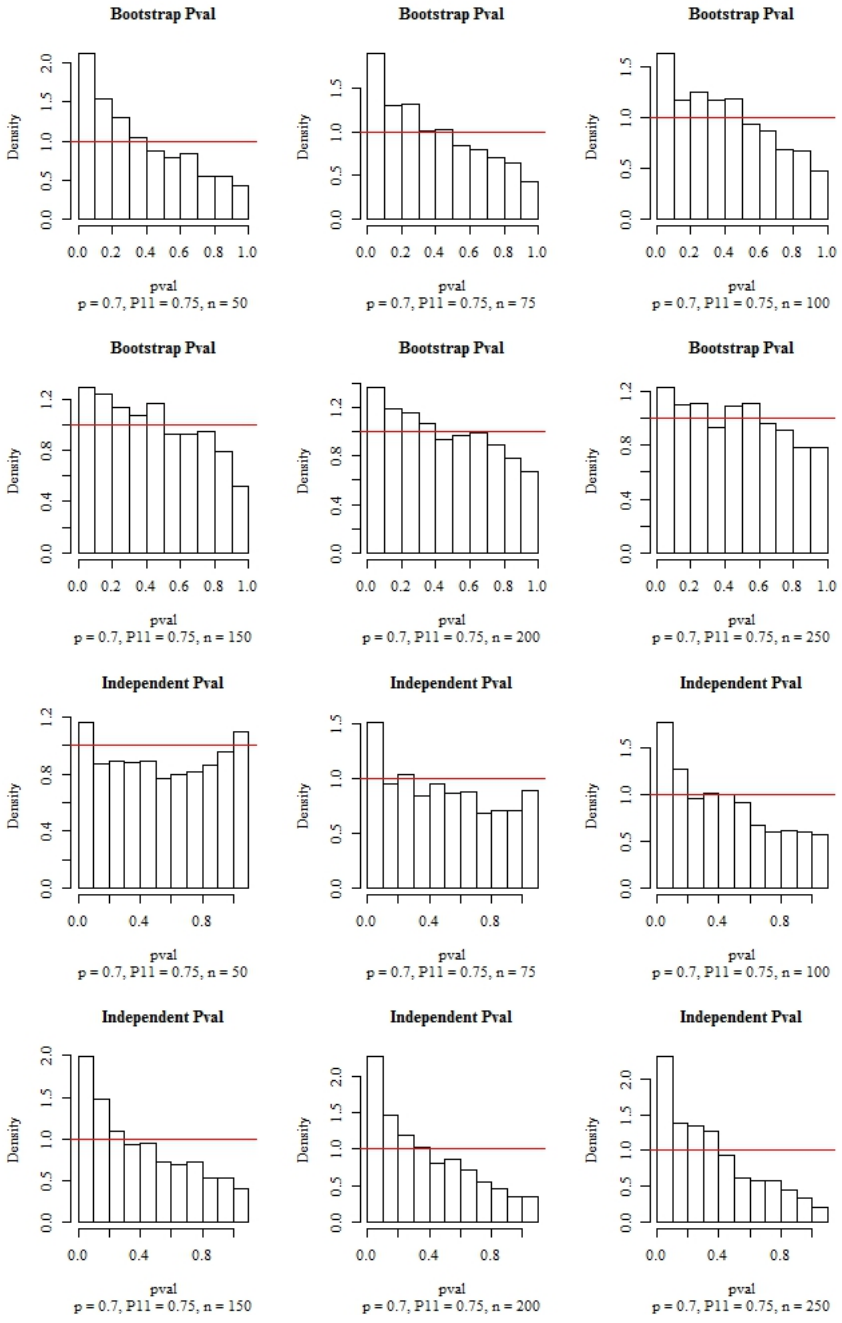


Figure 4.10: Histograms of p-values for G^2 Bootstrap and Independent p-values for Simulation (M) with Varying Sample Sizes

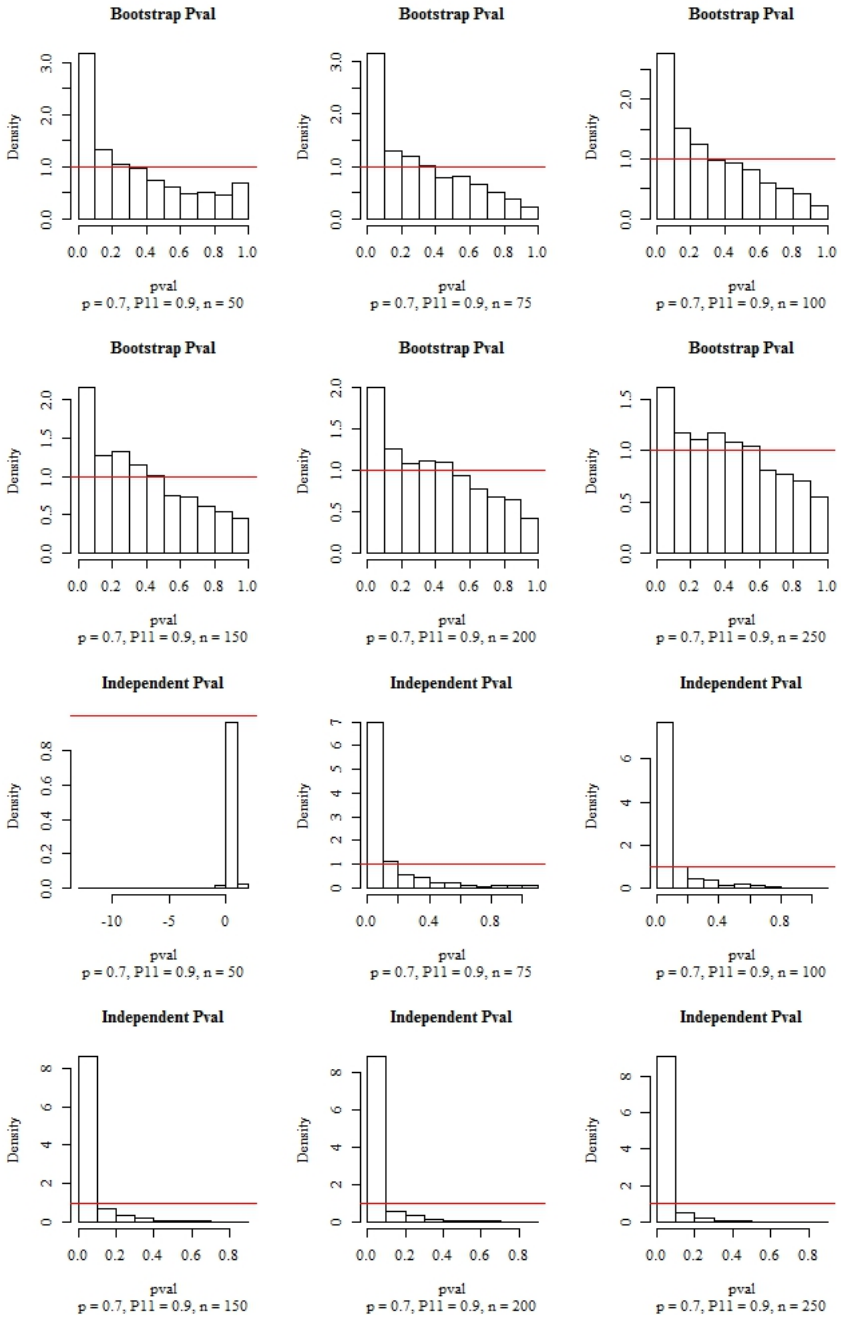


Table 4.14: APC of DCUSUM and Dependent LRT

n		Brownian DCUSUM	Worsley DCUSUM	Bootstrap LRT
$\tau = (1/5)n$				
75	M_p	0.05	0.11	0.29
	$S_{p,P_{11}}$	0.03	0.04	0.14
150	M_p	0.05	0.10	0.29
	$S_{p,P_{11}}$	0.10	0.11	0.18
300	M_p	0.12	0.18	0.44
	$S_{p,P_{11}}$	0.32	0.28	0.45
$\tau = (2/5)n$				
75	M_p	0.02	0.10	0.36
	$S_{p,P_{11}}$	0.06	0.10	0.26
150	M_p	0.11	0.20	0.40
	$S_{p,P_{11}}$	0.19	0.21	0.37
300	M_p	0.40	0.48	0.66
	$S_{p,P_{11}}$	0.46	0.43	0.69
$\tau = (1/2)n$				
75	M_p	0.03	0.11	0.40
	$S_{p,P_{11}}$	0.07	0.12	0.30
150	M_p	0.19	0.29	0.43
	$S_{p,P_{11}}$	0.22	0.23	0.41
300	M_p	0.54	0.60	0.66
	$S_{p,P_{11}}$	0.48	0.45	0.71

the theoretical power by comparing the test statistics $T^2_{\max}$ and $G^2_{\max}$ under a specified null model to the statistics generated under various alternative models.

The APC comparison can be found in Table 4.14. The models (M_p) and ($S_{p,P_{11}}$) are used to simulate various differences of the parameters $p(1)$ and $p(2)$ as well as $P_{11}(1)$ and $P_{11}(2)$. Descriptions of these model parameters can be found in Table 4.2. The bootstrap LRT outperforms both of the DCUSUM procedures in each model for all of the change point locations. The DCUSUM procedures show a much larger increase in power as the change point location moves towards the center of the sequence. All methods show an improvement in power when the change point is closer to the center of the sequence.

In order to compare the theoretical power of the $T^2_{\max}$ and $G^2_{\max}$ statistics, a null model must be assumed. This is due to the facts that when H_0 is false, there is no true null distribution and the null distribution depends on the parameters p and P_{11}. Because of this,

a rather arbitrary choice of p and P_{11} is used to calculate the critical value of a level $\alpha = 0.05$ test to compare the power of the two statistics. The choice of parameters under the null model are defined to be $p = 0.8$ and $P_{11} = 0.9$ to mimic a strong one step dependence with a moderate size of m.

The powers of the DCUSUM statistic $T^2_{\max}$ and dependent LRT statistic $G^2_{\max}$ for all models in Table 4.2 are calculated as follows. First, the 95^{th} percentiles of $T^2_{\max}$ and $G^2_{\max}$ under the null model (with parameters defined in the previous paragraph) are calculated from 2000 simulated values. Another 2000 test statistics are generated for each of the sample sizes using the alternative parameters for each of the five models. The power is then defined as the proportion of the test statistic values under the alternative model that are greater than or equal to the 95^{th} percentile under the null model.

The percentiles under the null hypothesis can be found in Table 4.15. The powers for each model and sample size are contained in Tables 4.16, 4.17, and 4.18.

The dependent LRT is a more powerful test for a majority of the models. Excluding the models where $\tau = (1/5)n$ and $n \leq 100$, the dependent LRT has higher power than DCUSUM procedures. Even for moderate sample sizes ($150 \leq n \leq 300$), the dependent LRT greatly outperforms the DCUSUM method. This is not consistent with the results under the assumption of independence stated in the literature. For example, see Robbins et al. [34], where it is noted that the independent CUSUM test has higher power than the independent LRT near the center of the data. One possible reason for this difference is that both of the statistics $T^2_{\max}$ and $G^2_{\max}$ are dependent on the parameter P_{11}, even for moderate sample sizes. For large samples ($n \geq 500$) and large differences in parameter values (models (L_p) and ($\mathrm{M}_{p,P_{11}}$)), the power of the two methods is comparable.

As expected, the power of the DCUSUM method suffers when the change point is closer to the edge of the data set ($\tau = (1/5)n$). Excluding the case of (L_p) for $n \leq 100$, the power increases substantially for each change point location shift closer to the midpoint of the data set. The power of the dependent LRT is also lower at the edge of the data set ($\tau = (1/5)n$),

Table 4.15: 95^{th} Percentiles of $T^2_{\max}$ and $G^2_{\max}$ under $H_0 : p = 0.8, P_{11} = 0.9$

n	$T^2_{\max}$	$G^2_{\max}$
50	10.9108	7.6000
75	9.7816	8.9286
100	10.0200	9.9688
150	9.7662	10.4294
200	9.7808	11.0691
250	9.7256	11.3444
300	9.2549	11.6931
500	9.4757	12.7741
1000	9.4524	12.5742

but the powers when $\tau = (2/5)n$ and $\tau = (1/2)n$ are similar. Both methods show improved power as the change point location gets closer to the midpoint of the data set.

From a practical power perspective, the dependent LRT is recommended for small to moderate sample sizes, or if the change point location is assumed to be far from the midpoint of the data. Due to the long run time of the bootstrap procedure, DCUSUM procedures are recommended for large samples, unless a small change in p is to be detected. This is not an issue because the powers of the two methods are comparable when $n \approx 1000$ and the change in p is moderate.

Table 4.16: Empirical Power Comparison of DCUSUM and Dependent LRT, $\tau = (1/5)n$

n		DCUSUM	Dependent LRT
50	L_p	0.23	0.11
	M_p	0.06	0.10
	S_p	0.01	0.08
	$M_{p,P_{11}}$	0.04	0.26
	$S_{p,P_{11}}$	0.00	0.11
75	L_p	0.26	0.20
	M_p	0.05	0.17
	S_p	0.02	0.09
	$M_{p,P_{11}}$	0.22	0.40
	$S_{p,P_{11}}$	0.03	0.15
100	L_p	0.24	0.29
	M_p	0.03	0.19
	S_p	0.01	0.10
	$M_{p,P_{11}}$	0.37	0.52
	$S_{p,P_{11}}$	0.05	0.15
150	L_p	0.28	0.54
	M_p	0.05	0.31
	S_p	0.02	0.14
	$M_{p,P_{11}}$	0.63	0.79
	$S_{p,P_{11}}$	0.11	0.29
200	L_p	0.34	0.68
	M_p	0.07	0.38
	S_p	0.02	0.15
	$M_{p,P_{11}}$	0.78	0.90
	$S_{p,P_{11}}$	0.18	0.37
250	L_p	0.40	0.79
	M_p	0.08	0.48
	S_p	0.03	0.18
	$M_{p,P_{11}}$	0.87	0.96
	$S_{p,P_{11}}$	0.28	0.47
300	L_p	0.51	0.88
	M_p	0.16	0.57
	S_p	0.06	0.23
	$M_{p,P_{11}}$	0.94	0.98
	$S_{p,P_{11}}$	0.37	0.57
500	L_p	0.79	0.99
	M_p	0.30	0.84
	S_p	0.09	0.33
	$M_{p,P_{11}}$	0.99	1.00
	$S_{p,P_{11}}$	0.60	0.81
1000	L_p	1.00	1.00
	M_p	0.84	1.00
	S_p	0.26	0.76
	$M_{p,P_{11}}$	1.00	1.00
	$S_{p,P_{11}}$	0.90	0.99

Table 4.17: Empirical Power Comparison of DCUSUM and Dependent LRT, $\tau = (2/5)n$

n		DCUSUM	Dependent LRT
50	L_p	0.01	0.19
	M_p	0.02	0.16
	S_p	0.03	0.09
	$M_{p,P_{11}}$	0.04	0.50
	$S_{p,P_{11}}$	0.01	0.21
75	L_p	0.07	0.39
	M_p	0.03	0.27
	S_p	0.02	0.12
	$M_{p,P_{11}}$	0.32	0.71
	$S_{p,P_{11}}$	0.07	0.27
100	L_p	0.12	0.55
	M_p	0.04	0.36
	S_p	0.03	0.14
	$M_{p,P_{11}}$	0.47	0.81
	$S_{p,P_{11}}$	0.10	0.37
150	L_p	0.34	0.82
	M_p	0.12	0.55
	S_p	0.05	0.22
	$M_{p,P_{11}}$	0.76	0.94
	$S_{p,P_{11}}$	0.20	0.54
200	L_p	0.52	0.92
	M_p	0.22	0.66
	S_p	0.09	0.29
	$M_{p,P_{11}}$	0.89	0.98
	$S_{p,P_{11}}$	0.31	0.63
250	L_p	0.66	0.98
	M_p	0.31	0.79
	S_p	0.10	0.36
	$M_{p,P_{11}}$	0.96	1.00
	$S_{p,P_{11}}$	0.39	0.73
300	L_p	0.80	0.99
	M_p	0.46	0.87
	S_p	0.17	0.38
	$M_{p,P_{11}}$	0.99	1.00
	$S_{p,P_{11}}$	0.52	0.80
500	L_p	0.98	1.00
	M_p	0.78	0.98
	S_p	0.31	0.60
	$M_{p,P_{11}}$	1.00	1.00
	$S_{p,P_{11}}$	0.76	0.96
1000	L_p	1.00	1.00
	M_p	0.99	1.00
	S_p	0.65	0.95
	$M_{p,P_{11}}$	1.00	1.00
	$S_{p,P_{11}}$	0.97	1.00

Table 4.18: Empirical Power Comparison of DCUSUM and Dependent LRT, $\tau = (1/2)n$

n		DCUSUM	Dependent LRT
50	L_p	0.01	0.21
	M_p	0.01	0.16
	S_p	0.02	0.12
	$M_{p,P_{11}}$	0.04	0.56
	$S_{p,P_{11}}$	0.02	0.23
75	L_p	0.05	0.42
	M_p	0.03	0.29
	S_p	0.03	0.14
	$M_{p,P_{11}}$	0.29	0.74
	$S_{p,P_{11}}$	0.08	0.34
100	L_p	0.13	0.62
	M_p	0.06	0.39
	S_p	0.05	0.15
	$M_{p,P_{11}}$	0.47	0.83
	$S_{p,P_{11}}$	0.11	0.38
150	L_p	0.42	0.86
	M_p	0.18	0.60
	S_p	0.09	0.25
	$M_{p,P_{11}}$	0.74	0.95
	$S_{p,P_{11}}$	0.22	0.55
200	L_p	0.64	0.95
	M_p	0.32	0.73
	S_p	0.13	0.30
	$M_{p,P_{11}}$	0.89	0.99
	$S_{p,P_{11}}$	0.31	0.66
250	L_p	0.78	0.98
	M_p	0.44	0.84
	S_p	0.18	0.36
	$M_{p,P_{11}}$	0.96	1.00
	$S_{p,P_{11}}$	0.41	0.76
300	L_p	0.88	1.00
	M_p	0.60	0.89
	S_p	0.24	0.44
	$M_{p,P_{11}}$	0.99	1.00
	$S_{p,P_{11}}$	0.53	0.84
500	L_p	0.99	1.00
	M_p	0.86	0.99
	S_p	0.40	0.64
	$M_{p,P_{11}}$	1.00	1.00
	$S_{p,P_{11}}$	0.77	0.96
1000	L_p	1.00	1.00
	M_p	0.99	1.00
	S_p	0.74	0.96
	$M_{p,P_{11}}$	1.00	1.00
	$S_{p,P_{11}}$	0.97	1.00

Chapter 5

Proposed Multipath and Multinomial Methods with Motivating Application

The main results of Chapters 2 and 3 provide single path methods to detect a change point in an m-dependent sequence of Bernoulli random variables. A multipath approach may provide a better test when several sequences $\{y_i\}_{i=1}^s$ are available. Recall that a *multipath* procedure will use the information from all of the $i = 1, 2, \ldots, s$ sequences y_i. The aim of this chapter is to explore multipath methods, including a proposed maximal change count statistic, and generalize the tests for Bernoulli sequences to the multinomial case. An application of these methods that motivated the one step Markov dependence assumption is provided at the conclusion of this chapter.

5.1 Maximal Change Count Statistic $\Delta_{\max}$

Define the column vector containing the values of each of the s sequences $\{y_i\}_{i=1}^s$ at time t to be X_t. When each sequence is made up of Bernoulli random variables, each entry x_{it} of X_t is a binary entry. The parameters corresponding to x_{it} are $p_i = p_{it} = \Pr(x_{it} = 1)$ and are the same for all values of t. An example of such a vector at time t with $s = 4$ four sequences

$\{y_i\}_{i=1}^4$ is:

$$X_t = \begin{pmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \\ x_{4t} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}.$$

For each time t, all pairwise differences of vectors $\{X_j\}_{j=1}^t$ up to time t will be computed and the norms of the resulting difference vectors will be calculated. The difference vectors are defined as $D_{rq,t} = X_r - X_q$ for $1 \leq q < r \leq t$ for each $1 \leq t \leq n$, with entries $d_{i,rq} = x_{ir} - x_{iq}$. The sets of pairwise differences $\{D_{rq,t}\}_{1 \leq q < r \leq t}$ for each value of t will be used to construct the test statistic.

When no change is present in the sequences, the difference vectors will consist of mostly zero entries. A comparison of all difference vectors $D_{rq,t}$ for $1 \leq q < r \leq t$ gives information about the movement of the sequences to and from states. If the vector $D_{rq,t}$ has a large number of nonzero entries, this gives evidence of an abrupt change in parameters, and hence, evidence of a change point τ. This statistic is not only capable of estimating and detecting a change point location, but it also provides insight on which sequences are contributing to the change in parameter values.

The detection criteria requires that a norm be used to measure the maximum value of $D_{rq,t}$ for each time t. The norm used for the proposed method is the Frobenius norm:

$$||D_{rq,t}||_{F_1} = \sum_{i=1}^{s} |d_{i,rq}|, \tag{5.1}$$

where $|\cdot|$ denotes Euclidean distance. The norms of each pair of r and q is calculated and, for each t, the result is a random field of $(t-1)t/2$ elements.

Define $\Delta_t = \max_{1 \leq q < r \leq t} ||D_{rq,t}||_{F_1}$, then the test statistic for change point detection is the *maximal change count statistic* $\Delta_{\max} = \max_{1 < t < n} \Delta_t$. The aim of the maximal change count statistic is to provide a multipath method that is able to both detect a change point and

give information on the sequences that contribute to the change.

5.1.1 Small Sample Distribution of $||D_{rq}||_{F_1}$

Suppose that each sequence $y_i \sim$ Bernoulli(p_i). Define $Y = (y_1, y_2, \ldots, y_s)'$, $\boldsymbol{p} = (p_1, p_2, \ldots, p_s)'$, and $\boldsymbol{P}$ as in (1.18). The hypotheses used to detect a common change point τ in the structure of Y, which are a generalization of those in Section 1.4.1 are:

H_0 : $x_{it} \sim$ Bernoulli(p_i) with transition probabilities $P_{uv,i}$ for all times t,

H_a : There exists τ, $1 < \tau < n$, such that

$x_{it} \sim$ Bernoulli$(p_i(1))$ for all $1 < t \leq \tau$ and $x_{it} \sim$ Bernoulli$(p_i(2))$ for all $\tau < t \leq n$

where $\boldsymbol{p}(1) \neq \boldsymbol{p}(2)$ and the events after the change are independent of the events prior to the change.

The simplest case of these hypotheses occurs when the entries in the probability vector $\boldsymbol{p}$ and the values of $P_{u,v,i,t}$ are the same for all i time series. Specifically, $P_{u,v,i,t} = P_{uv}$ for all $1 \leq i \leq s$ and $1 \leq t \leq n$. Under H_0, these assumptions lead to the same system of equations (1.20) with solution (1.21).

In order to perform hypothesis testing and inference about a potential change point $\hat{\tau}$, the distirbution of $||D_{rq}||_{F_1}$ must be explored. Before determining the distribution, a lemma describing the distribution of the entries $|d_{i,rq}|$ of $||D_{rq}||_{F_1}$ is required.

Lemma 5.1.1 *Let x_{ir} and x_{iq} be the i^{th} components of the corresponding random vectors X_r and X_q. Define the transition matrix $\boldsymbol{P}_t$ from time t to time $t+1$ as in equation* (1.18)*. If $x_{it} \sim$ Bernoulli(p) and $\boldsymbol{P}_t = \boldsymbol{P}$ for all times $t = 1, 2, \ldots, n$ and sequences $i = 1, 2, \ldots, s$, then the random variables $|d_{i,rq}| = |x_{ir} - x_{iq}|$ have pmf equal to:*

$$f_{|d_{i,rq}|}(z) = \begin{cases} \boldsymbol{P}_{00}^{|r-q|}(1-p) + \boldsymbol{P}_{11}^{|r-q|}p & \text{if } z = 0, \\ \boldsymbol{P}_{01}^{|r-q|}(1-p) + \boldsymbol{P}_{10}^{|r-q|}p & \text{if } z = 1. \end{cases}$$

Proof By definition, the random variables $|d_{i,rq}|$ must take on the values 0 or 1. The probabilities are calculated directly:

$$\begin{aligned}
\Pr(|d_{i,rq}| = 0) &= \Pr(x_{iq} = 0 \cap x_{ir} = 0) + \Pr(x_{iq} = 1 \cap x_{ir} = 1) \\
&= \Pr(x_{ir} = 0 \mid x_{iq} = 0)\Pr(x_{iq} = 0) \\
&\quad + \Pr(x_{ir} = 1 \mid x_{iq} = 1)\Pr(x_{iq} = 1) \\
&= \boldsymbol{P}_{00}^{|r-q|}(1-p) + \boldsymbol{P}_{11}^{|r-q|}p, \qquad (5.2) \\
\Pr(|d_{i,rq}| = 1) &= \Pr(x_{iq} = 0 \cap x_{ir} = 1) + \Pr(x_{iq} = 0 \cap x_{ir} = 1) \\
&= \Pr(x_{ir} = 1 \mid x_{iq} = 0)\Pr(x_{iq} = 0) \\
&\quad + \Pr(x_{i,r} = 0 \mid x_{iq} = 1)\Pr(x_{ir} = 1) \\
&= \boldsymbol{P}_{01}^{|r-q|}(1-p) + \boldsymbol{P}_{10}^{|r-q|}p. \qquad (5.3)
\end{aligned}$$

Equations (5.2) and (5.3) follow from basic properties of the transition matrix $\boldsymbol{P}$. The *pmf* of $|d_{i,rq}|$ is given as:

$$f_{|d_{i,rq}|}(z) = \begin{cases} \boldsymbol{P}_{00}^{|r-q|}(1-p) + \boldsymbol{P}_{11}^{|r-q|}p, & \text{if } z = 0, \\ \boldsymbol{P}_{01}^{|r-q|}(1-p) + \boldsymbol{P}_{10}^{|r-q|}p & \text{if } z = 1. \end{cases}$$

∎

With this lemma in hand, the general distribution of $||D_{rq}||_{F_1}$ can be stated for any values of $1 \leq q < r \leq n$.

Theorem 5.1.2 *Assume that $x_{it} \sim$ Bernoulli(p) and $\boldsymbol{P}_t = \boldsymbol{P}$ for all times $t = 1, 2, \ldots, n$ and sequences $i = 1, 2, \ldots, s$, and that the rows of X_t are independent for all t. Under the hypothesis of no change, the components of X_t are independent and identically distributed*

Bernoulli(p) random variables with pmf:

$$f_{m_{it}}(x) = \begin{cases} p & \text{if } x = 1, \\ 1-p & \text{if } x = 0, \end{cases}$$

and $||D_{rq}||_{F_1} \sim Binom(s, \boldsymbol{P}_{01}^{|r-q|}(1-p) + \boldsymbol{P}_{10}^{|r-q|}p)$.

Proof First, the distribution of $|d_{i,rq}|$ must be found. By definition, $|d_{i,rq}| = |x_{ir} - x_{iq}|$ where x_{ir} and x_{iq} are dependent Bernoulli(p) random variables with dependence given by the values of $\boldsymbol{P}^{|r-q|}$. From Lemma 5.1.1, the *pmf* of $|d_{i,rq}|$ is:

$$f_{|d_{i,rq}|}(z) = \begin{cases} \boldsymbol{P}_{00}^{|r-q|}(1-p) + \boldsymbol{P}_{11}^{|r-q|}p, & \text{if } z = 0, \\ \boldsymbol{P}_{01}^{|r-q|}(1-p) + \boldsymbol{P}_{10}^{|r-q|}p & \text{if } z = 1. \end{cases}$$

With these probabilities in hand, the *mgf* of $|d_{i,rq}|$ can be computed. Let $\tilde{p} = \boldsymbol{P}_{01}^{|r-q|}(1-p) + \boldsymbol{P}_{10}^{|r-q|}p$, then:

$$M_{|d_{i,rq}|}(t^*) = \mathrm{E}e^{t^*|d_{i,rq}|} = (1-\tilde{p}) + \tilde{p}e^{t^*}.$$

Let $A \sim$ Bernoulli($\tilde{p}$), then $M_{|d_{i,rq}|}(t^*) = M_A(t^*)$ for all t^*. By Theorem 2.3.11b in Casella and Berger [3], $|d_{i,rq}| \sim$ Bernoulli($\tilde{p}$). Therefore:

$$f_{|d_{i,rq}|}(z) = \begin{cases} 1-\tilde{p} & \text{if } z = 0, \\ \tilde{p} & \text{if } z = 1. \end{cases}$$

The sequences y_i, and hence, the rows of X_t are independent for all values of t. Therefore, the rows of D_{rq} are also independent. By the assumption that $p_i = p$ for all i and the values of $\boldsymbol{P}$ do not depend on the time t, the random variables $|d_{i,rq}|$ are independent for all i. The sum of n_d *iid* $|d_{i,rq}|$ random variables has *mgf*:

$$M_{\sum |d_{i,rq}|}(t^*) = [M_{|d_{i,rq}|}(t^*)]^{n_d} = [(1-\tilde{p}) + \tilde{p}e^{t^*}]^{n_d}.$$

Let $B \sim \text{Binomial}(n_d, \tilde{p})$, then $M_{|d_{i,rq}|}(t^*) = M_B(t^*)$ for all t^*. Again, by Theorem 2.3.11b in Casella and Berger [3], $|d_{i,rq}| \sim \text{Binomial}(n_d, \tilde{p})$. Therefore, $||D_{rq}||_{F_1} \sim \text{Binom}(n_d, \tilde{p})$.

To finish the proof, the value of n_d must be determined. Each random vector X_t is composed of s random variables. Therefore, D_{rq} is the sum of s *iid* random variables. Hence, $n_d = s$. ∎

The result in Theorem 5.1.2 is a restricted case that may not be applicable in most settings. The most general case is to consider sequences with unique success probabilities p_i as well as unique transition matrices $\boldsymbol{P}_i$. In this case, the resulting distribution of $||D_{rq}||_{F_1}$ is a Poisson-Binomial distribution, which is believed to be studied first by S. Poisson in 1837. A summary of results for the Poisson-Binomial distribution is given by Wang [37]. This distribution is a generalization of the binomial distribution where the independent Bernoulli trials are able to have different success probabilities.

Let $B_i \sim \text{Bernoulli}(p_i)$, then the random variable $B = \sum B_i$ is a Poisson-Binomial distribution with mean $\mu = \sum p_i$ and variance $\sigma^2 = \sum(1 - p_i)p_i$. The following theorem describes the distribution of $||D_{rq}||_{F_1}$ under the most general assumptions.

Theorem 5.1.3 *Assume that $x_{it} \sim$ Bernoulli(p_i) with transition matrix $\boldsymbol{P}_i$ for all times $t = 1, 2, \ldots, n$, and that the rows of X_t are independent for all t. Under the hypothesis of no change, the entries of X_t are independent and identically distributed Bernoulli(p_i) random variables for all time points $1 \leq t \leq n$. That is, for all i and t:*

$$f_{x_{it}}(z) = \begin{cases} p_i & \text{if } z = 1, \\ 1 - p_i & \text{if } z = 0. \end{cases}$$

Then $||D_{rq}||_{F_1}$ follows a Poisson-Binomial distribution with parameters:

$$\tilde{p}_{i,rq} = \boldsymbol{P}_{01,i}^{|r-q|}(1 - p_i) + \boldsymbol{P}_{10,i}^{|r-q|} p_i \text{ for } 1 \leq i \leq s.$$

Proof Following similar arguments in the proof of Theorem 5.1.2, it can be shown that:

$$f_{|d_{i,rq}|}(x) = \begin{cases} 1 - \tilde{p}_{i,rq} & \text{if } x = 0, \\ \tilde{p}_{i,rq} & \text{if } x = 1, \end{cases}$$

and hence, $|d_{i,rq}| \sim \text{Bernoulli}(\tilde{p}_{i,rq})$. Since $||D_{rq}||_{F_1} = \sum |d_{i,rq}|$, it follows by definition that $||D_{rq}||_{F_1}$ has a Poisson-Binomial distribution with parameters $\tilde{p}_{i,rq}$ for $1 \leq i \leq s$. ∎

5.1.2 Covariance structure of $||D_{rq}||_{F_1}$

It is clear from the construction of the matrices D_{rq} that for two sets of indices $\{r, q\}$ and $\{r', q'\}$ the F_1 norms of D_{rq} and $D_{r'q'}$ have a nontrivial covariance structure. The goal of this section is to explore that structure to aid in determining the asymptotic distribution of Δ_t. For the remainder of this section, it is assumed that $\boldsymbol{P}_{i,t} = \boldsymbol{P}$ and $p_i = p$ for all i and t. That is, the success probabilities and transition probabilities are equal for all times and sequences.

The indices of D_{rq} must follow a specific ordering. For a given time t, the set $\{q, q', r, r'\}$ must satisfy $1 \leq q < r \leq t$ and $1 \leq q' < r' \leq t$. With these restrictions, there are exactly seven cases for ordering of the indices. The partial covariance for all seven cases in terms of the entries of $\boldsymbol{P}$ may be found in Table 5.1. To obtain the full covariance, subtract the value of $\text{E}(||d_{rq}||)\text{E}(||d_{r'q'}||)$ from each as defined in equation (5.5).

For illustration purposes, the calculations for Case I are provided below. The other cases are similar.

$$\begin{aligned} \text{E}(||d_{rq}||\,||d_{r'q'}||) &= \Pr(||d_{rq}|| = 1 \cap ||d_{r'q'}|| = 1) \\ &= \Pr(x_r = 0 \mid x_{q'} = 1)\Pr(x_{q'} = 1 \mid x_q = 1)\Pr(x_q = 1) \\ &\quad + \Pr(x_r = 1 \mid \cap x_{q'} = 0)\Pr(x_{q'} = 0 \mid x_q \ = 0)\Pr(x_q \ = 0) \\ &= \boldsymbol{P}_{10}^{|r'-q|}\boldsymbol{P}_{11}^{|q'-q|}p + \boldsymbol{P}_{01}^{|r-q'|}\boldsymbol{P}_{00}^{|q'-q|}(1-p). \end{aligned} \tag{5.4}$$

Table 5.1: Partial Covariance for all 7 Cases in the Bernoulli Setting

Case	Order of Indices	$\mathrm{E}(\|\|D_{rq}\|\|_{F_1}\|\|D_{r'q'}\|\|_{F_1})$
I	$1 \le q < q' < r = r' \le n$	$\boldsymbol{P}_{10}^{\|r-q'\|}\boldsymbol{P}_{11}^{\|q'-q\|}p + \boldsymbol{P}_{01}^{\|r-q'\|}\boldsymbol{P}_{00}^{\|q'-q\|}(1-p)$
II	$1 \le q = q' < r < r' \le n$	$\boldsymbol{P}_{00}^{\|r'-r\|}\boldsymbol{P}_{10}^{\|r-q\|}p + \boldsymbol{P}_{11}^{\|r'-r\|}\boldsymbol{P}_{01}^{\|r-q\|}(1-p)$
III	$1 \le q < r < q' < r' \le n$	$\boldsymbol{P}_{10}^{\|r'-q'\|}\boldsymbol{P}_{01}^{\|q'-r\|}\boldsymbol{P}_{10}^{\|r-q\|}p + \boldsymbol{P}_{01}^{\|r'-q'\|}\boldsymbol{P}_{00}^{\|q'-r\|}\boldsymbol{P}_{10}^{\|r-q\|}p$ $+\boldsymbol{P}_{01}^{\|r'-q'\|}\boldsymbol{P}_{11}^{\|q'-r\|}\boldsymbol{P}_{01}^{\|r-q\|}(1-p) + \boldsymbol{P}_{01}^{\|r'-q'\|}\boldsymbol{P}_{10}^{\|q'-r\|}\boldsymbol{P}_{01}^{\|r-q\|}(1-p)$
IV	$1 \le q < r = q' < r' \le n$	$\boldsymbol{P}_{01}^{\|r'-r\|}\boldsymbol{P}_{10}^{\|r-q\|}p + \boldsymbol{P}_{10}^{\|r'-r\|}\boldsymbol{P}_{01}^{\|r-q\|}(1-p)$
V	$1 \le q < q' < r < r' \le n$	$\boldsymbol{P}_{00}^{\|r'-r\|}\boldsymbol{P}_{10}^{\|r-q'\|}\boldsymbol{P}_{11}^{\|q'-q\|}p + \boldsymbol{P}_{01}^{\|r'-r\|}\boldsymbol{P}_{00}^{\|r-q'\|}\boldsymbol{P}_{10}^{\|q'-q\|}p$ $+\boldsymbol{P}_{10}^{\|r'-r\|}\boldsymbol{P}_{11}^{\|r-q'\|}\boldsymbol{P}_{01}^{\|q'-q\|}(1-p) + \boldsymbol{P}_{11}^{\|r'-r\|}\boldsymbol{P}_{01}^{\|r-q'\|}\boldsymbol{P}_{00}^{\|q'-q\|}(1-p)$
VI	$1 \le q < q' < r' < r \le n$	$\boldsymbol{P}_{00}^{\|r-r'\|}\boldsymbol{P}_{10}^{\|r'-q'\|}\boldsymbol{P}_{11}^{\|q'-q\|}p + \boldsymbol{P}_{10}^{\|r-r'\|}\boldsymbol{P}_{01}^{\|r'-q'\|}\boldsymbol{P}_{10}^{\|q'-q\|}p$ $+\boldsymbol{P}_{01}^{\|r-r'\|}\boldsymbol{P}_{10}^{\|r'-q'\|}\boldsymbol{P}_{01}^{\|q'-q\|}(1-p) + \boldsymbol{P}_{11}^{\|r-r'\|}\boldsymbol{P}_{01}^{\|r'-q'\|}\boldsymbol{P}_{00}^{\|q'-q\|}(1-p)$
VII	$1 \le q' < q < r < r' \le n$	$\boldsymbol{P}_{00}^{\|r'-r\|}\boldsymbol{P}_{10}^{\|r-q\|}\boldsymbol{P}_{11}^{\|q-q'\|}p + \boldsymbol{P}_{10}^{\|r'-r\|}\boldsymbol{P}_{01}^{\|r-q\|}\boldsymbol{P}_{10}^{\|q-q'\|}p$ $+\boldsymbol{P}_{01}^{\|r'-r\|}\boldsymbol{P}_{10}^{\|r-q\|}\boldsymbol{P}_{01}^{\|q-q'\|}(1-p) + \boldsymbol{P}_{11}^{\|r'-r\|}\boldsymbol{P}_{01}^{\|r-q\|}\boldsymbol{P}_{00}^{\|q-q'\|}(1-p)$

The expected value of any single random variable $||d_{rq}||$ is:

$$\begin{aligned}
\mathrm{E}(||d_{rq}||) &= \Pr(||d_{rq}|| = 1) \\
&= \Pr(x_r = 0 \cap x_q = 1) + \Pr(x_r = 1 \cap x_q = 0) \\
&= \Pr(x_r = 0 \mid x_q = 1)\Pr(x_q = 1) + \Pr(x_r = 1 \mid x_q = 0)\Pr(x_q = 0) \\
&= \boldsymbol{P}_{10}^{|r-q|}p + \boldsymbol{P}_{01}^{|r-q|}(1-p). \qquad (5.5)
\end{aligned}$$

Combining equations (5.4) and (5.5) yields the covariance of any two variables $||d_{rq}||$ and $||d_{r'q'}||$ in Case I:

$$\begin{aligned}
\mathrm{Cov}(||d_{rq}||\,||d_{rq'}||) &= \mathrm{E}(||d_{rq}||\,||d_{rq'}||) - \mathrm{E}(||d_{rq}||)\mathrm{E}(||d_{rq'}||) \\
&= \boldsymbol{P}_{10}^{|r-q'|}\boldsymbol{P}_{11}^{|q'-q|}p + \boldsymbol{P}_{01}^{|r-q'|}\boldsymbol{P}_{00}^{|q'-q|}(1-p) \\
&\quad - \left(\boldsymbol{P}_{10}^{|r-q|}p + \boldsymbol{P}_{01}^{|r-q|}(1-p)\right)\left(\boldsymbol{P}_{10}^{|r-q'|}p + \boldsymbol{P}_{01}^{|r-q'|}(1-p)\right). \qquad (5.6)
\end{aligned}$$

Combining (5.6) and the fact that rows of X_t are independent gives the covariance of two

difference matrices in Case I:

$$\begin{aligned}
\text{Cov}(||D_{rq}||\,||D_{rq'}||) &= \text{Cov}\left(\sum_{i=1}^{s}||d_{i,rq}||\,\sum_{j=1}^{s}||d_{j,rq'}||\right)\\
&= \sum_{i=1}^{s}\sum_{j=1}^{s}\text{Cov}\left(||d_{i,rq}||,||d_{j,rq'}||\right)\\
&= \sum_{i=1}^{s}\text{Cov}\left(||d_{i,rq}||,||d_{i,rq'}||\right)\\
&= s\text{Cov}\left(||d_{rq}||,||d_{rq'}||\right)\\
&= s\Big[\boldsymbol{P}_{10}^{|r-q'|}\boldsymbol{P}_{11}^{|q'-q|}p+\boldsymbol{P}_{01}^{|r-q'|}\boldsymbol{P}_{00}^{|q'-q|}(1-p)\\
&\quad-\left(\boldsymbol{P}_{10}^{|r-q|}p+\boldsymbol{P}_{01}^{|r-q|}(1-p)\right)\left(\boldsymbol{P}_{10}^{|r-q'|}p+\boldsymbol{P}_{01}^{|r-q'|}(1-p)\right)\Big].
\end{aligned}$$

5.1.3 Change point Detection with $||D_{rq}||_{F_1}$

Under the assumptions of Theorem 5.1.2, $||D_{rq}||_{F_1}$ follows a binomial distribution with parameters s and $\tilde{p}$. The parameter $\tilde{p}=\boldsymbol{P}_{01}^{|r-q|}(1-p)+\boldsymbol{P}_{10}^{|r-q|}p$ depends on the values of r and q. To complicate matters further, the covariance of $||D_{rq}||_{F_1}$ and $||D_{r'q'}||_{F_1}$ is dependent on the arrangement of the indices r, q, r', and q'. All of these facts cause difficulty in obtaining the asymptotic distribution of $\Delta_t=\max\limits_{1<q<r\leq t}||D_{rq}||_{F_1}$.

A two dimensional *random field* on the positive integers $\mathbb{Z}^2_{>0}$ is a collection of random variables with similar properties indexed by elements of $\mathbb{Z}^2_{>0}$. The dual indexing of D_{rq} allows these variables to be thought of as a random field of correlated binomial random variables. When large samples are available, each D_{rq} may be approximated by normal random variables, creating a correlated normal random field.

There are methods in the literature that provide asymptotic distributions for the maximum or minimum of random fields under certain conditions. One such approach is to consider an application of the Extremal Types Theorem described by Pereira [30], which provides asymptotic results for the distribution of the maximum of correlated normal variables in a random field. After careful exploration of the limiting covariance structure, this

route is not possible.

In order for the maximum of a correlated random field in two dimensions to have an asymptotic distribution via the Extremal Types Theorem, disjoint sub rectangles of the field must satisfy a limiting independence argument. The limiting property of the covariance is defined as $D(u_{\boldsymbol{n},\boldsymbol{i}})$ in Periera [30] and is restated below. If $\mathcal{F} \subset \mathbb{Z}^2_{>0}$ is a family of sets of indices, then there exist sequences of integer valued constants $\{k_{n_i}\}_{n_i \geq 1}$, $\{l_{n_i}\}_{n_i \geq 1}$, $i = 1, 2$, such that as $\boldsymbol{n} = (n_1, n_2) \to \boldsymbol{\infty}$, we have:

$$(k_{n_1}, k_{n_2}) \to \boldsymbol{\infty}, \left(\frac{k_{n_1} l_{n_1}}{n_1}, \frac{k_{n_2} l_{n_2}}{n_2}\right) \to \boldsymbol{\infty}, \text{ and } \left(k_{n_1}\Delta^{(1)}_{\boldsymbol{n},l_{n_1}}, k_{n_1}k_{n_2}\Delta^{(2)}_{\boldsymbol{n},l_{n_2}}\right) \to \mathbf{0}. \tag{5.7}$$

The values above may be interpreted in the following way. The level of separation between two sub rectangles of $\mathbb{Z}^2_{>0}$ is denoted as l_{n_i}, the values k_{n_i} are limiting constants, and the values of $\Delta^{(i)}_{\boldsymbol{n},l_{n_i}}$ are the components of the mixing coefficient, which is a measure of the dependence of elements in disjoint rectangles. If this requirement is met, then two disjoint sub rectangles may be thought of as nearly independent, and the Extremal Types Theorem may be applied to obtain an asymptotic distribution of $\Delta_{\max}$. Unfortunately, the covariance structure of $||D_{rq}||_{F_1}$ and $||D_{r'q'}||_{F_1}$ does not satisfy the third condition of (5.7). This is summarized below.

Remark The third assumption of the Extremal Types Theorem is violated by at least one case of the covariance structure of $||D_{rq}||_{F_1}$ and $||D_{r'q'}||_{F_1}$.

Proof Consider two disjoint rectangles in Case II, described in Table 5.1 and without loss of generality, suppose $p \in (0,1) \setminus \{\frac{1}{2}\}$. Fix the value q and r, and let $r' \to \infty$. Substituting the values given in Lemma 1.4.2 for $\lim_{t\to\infty} \boldsymbol{P}^t_{uv}$ when appropriate, the limit of the covariance is given as:

$$\begin{aligned}\lim_{r'\to\infty} \text{Cov}(||D_{rq}||_{F_1}, ||D_{r'q}||_{F_1}) = {} & p^2\boldsymbol{P}_{10}^{|r-q|} + (1-p)^2\boldsymbol{P}_{01}^{|r-q|} \\ & - \left[\boldsymbol{P}_{10}^{|r-q|}p + \boldsymbol{P}_{01}^{|r-q|}(1-p)\right][2p(1-p)]\end{aligned}$$

$$= (2p-1)\left[p^2\boldsymbol{P}_{10}^{|r-q|} - (1-p)^2\boldsymbol{P}_{01}^{|r-q|}\right]. \tag{5.8}$$

When $|r-q|$ is large, (5.8) is approximately equal to $-p(1-p)(2p-1) \neq 0$. Therefore, disjoint rectangles need not be independent, no matter how far apart they are. ∎

5.1.4 Restrictions and complications of $||D_{rq}||_{F_1}$

A known asymptotic distribution for Δ_t would provide a method to calculate p-values and aid in the use of the maximal change count statistic for change point detection. Consider the assumptions in Theorem 5.1.2. As n tends to infinity, the m-dependence assumption for each sequence y_i forces the parameters P_{uv} to approach p or $1-p$. Therefore, $||D_{rq}||_{F_1}$ is asymptotically Binom$(s, 2p(1-p))$.

The asymptotic distribution of the statistic $||D_{rq}||_{F_1}$ causes several problems in change point detection and inference. The success probability of the Binomial distribution depends on the parameter p, regardless of the sample size. Thus, neither the small sample nor asymptotic distribution of the statistic $||D_{rq}||_{F_1}$ is not pivotal, and must be modified in some way to be asymptotically independent of p.

Another issue is that under the alternative hypothesis, if $p(1) \approx 1-p(2)$, the asymptotic distributions before and after the change will have parameters $2p(1)(1-p(1))$ and $2(1-p(2))p(2)$, which are approximately equal. In this case, no change will be detected by the statistic Δ_t, even though a change in parameters has occurred.

An example of this situation is the large difference model (L) in Chapter 4, where $p(1) = 0.8$ and $p(2) = 0.2$. Both the DCUSUM procedure and dependent LRT are able to detect this change with reasonable power.

Future work includes identification of a function $g(\cdot)$ that will provide a pivotal quantity independent of p as well as exploration of the detection and estimation capabilities of the statistic Δ_t.

5.2 Multipath Dependent LRT

A natural generalization of the single path dependent LRT in Chapter 3 is the multipath dependent LRT. Suppose there are multiple sequences of Bernoulli random variables $\{y_i\}_{i=1}^{s}$ with parameters p_i, $\boldsymbol{P}_i$ and τ_i. The multipath dependent LRT will use the information from all s sequences to detect a change point in the sequences.

Consider the same hypotheses given in Section 5.1.1, except that under the alternative, the change points τ_i are allowed to vary for each sequence y_i. Define $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_s)'$ then:

H_0 : $x_{it} \sim$ Bernoulli(p_i) with transition probabilities $P_{uv,i}$ for all times t,

H_a : There exists $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_s)'$, $1 < \tau_i < n$, such that

$x_{it} \sim$ Bernoulli$(p_i(1))$ for all $1 < t \leq \tau_i$ and $x_{it} \sim$ Bernoulli$(p_i(2))$ for all $\tau_i < t \leq n$ where $\boldsymbol{p}(1) \neq \boldsymbol{p}(2)$ and the events after the change are independent of the events prior to the change.

The resulting likelihood functions for time $\boldsymbol{t} = (t_1, \ldots, t_s)'$ are:

$$
\begin{aligned}
L_{H_0}^{**} =& \prod_{i=1}^{s} P_{11,i}^{n_{11,i}^{n} - x_{t_i} x_{t_i+1}} P_{10,i}^{n_{10,i}^{n} - x_{t_i}(1-x_{t_i+1})} P_{01,i}^{n_{01,i}^{n} - (1-x_{t_i}) x_{t_i+1}} P_{00,i}^{n_{00,i}^{n} - (1-x_{t_i})(1-x_{t_i+1})}, \\
L_{H_a}^{*} =& \prod_{i=1}^{s} P_{00,i}^{n_{00,i}^{t_i}}(1) P_{01,i}^{n_{01,i}^{t_i}}(1) P_{10,i}^{n_{10,i}^{t_i}}(1) P_{11,i}^{n_{11,i}^{t_i}}(1) \\
& \times P_{11,i}(2)^{n_{11,i}^{n} - n_{11,i}^{t_i} - x_{t_i} x_{t_i+1}} P_{10,i}(2)^{n_{10,i}^{n} - n_{10,i}^{t_i} - x_{t_i}(1-x_{t_i+1})} \\
& \times P_{01,i}(2)^{n_{01,i}^{n} - n_{01,i}^{t_i} - (1-x_{t_i}) x_{t_i+1}} P_{00,i}(2)^{n_{00,i}^{n} - n_{00,i}^{t_i} - (1-x_{t_i})(1-x_{t_i+1})},
\end{aligned} \tag{5.9}
$$

with MLEs for $P_{00,i}$ and $P_{11,i}$:

$$
\begin{aligned}
\hat{P}_{00,i} &= \frac{n_{00,i}^{n} - (1-x_{t_i})(1-x_{t_i+1})}{n_{00,i}^{n} - (1-x_{t_i})(1-x_{t_i+1}) + n_{01,i}^{n} - (1-x_{t_i}) x_{t_i+1}}, \\
\hat{P}_{11,i} &= \frac{n_{11,i}^{n} - x_{t_i} x_{t_i+1}}{n_{11,i}^{n} - x_{t_i} x_{t_i+1} + n_{10,i}^{n} - x_{t_i}(1-x_{t_i+1})},
\end{aligned} \tag{5.10}
$$

and MLEs for $P_{00,i}(1), P_{11,i}(1), P_{00,i}(2),$ and $P_{11,i}(2)$:

$$\begin{aligned}
\hat{P}_{11,i} &= \frac{n^n_{11,i}}{n^n_{11,i} + n^n_{10,i}}, \quad \hat{P}_{00,i} = \frac{n^n_{00,i}}{n^n_{00,i} + n^n_{01,i}}, \\
\hat{P}_{11,i}(1) &= \frac{n^{t_i}_{11,i}}{n^{t_i}_{11,i} + n^{t_i}_{10,i}}, \quad \hat{P}_{00,i}(1) = \frac{n^{t_i}_{00,i}}{n^{t_i}_{00,i} + n^{t_i}_{01,i}}, \\
\hat{P}_{11,i}(2) &= \frac{n^n_{11,i} - n^{t_i}_{11,i} - x_{t_i}x_{t_i+1}}{n^n_{11,i} - n^{t_i}_{11,i} - x_{t_i}x_{t_i+1} + n^n_{10,i} - n^{t_i}_{10,i} - x_{t_i}(1 - x_{t_i+1})}, \\
\hat{P}_{00,i}(2) &= \frac{n^n_{00,i} - n^{t_i}_{00,i} - (1 - x_{t_i})(1 - x_{t_i+1})}{n^n_{00,i} - n^{t_i}_{00,i} - (1 - x_{t_i})(1 - x_{t_i+1}) + n^n_{01,i} - n^{t_i}_{01,i} - (1 - x_{t_i})x_{t_i+1}}.
\end{aligned} \tag{5.11}$$

Substituting these into the equations for $\boldsymbol{p}_0, \boldsymbol{p}_1$, and $\boldsymbol{p}_2$ gives the following MLEs:

$$\begin{aligned}
\hat{p}_{0,i} &= \frac{1 - \hat{P}_{00,i}}{2 - \hat{P}_{00,i} - \hat{P}_{11,i}}, \\
\hat{p}_{1,i} &= \frac{1 - \hat{P}_{00,i}(1)}{2 - \hat{P}_{00,i}(1) - \hat{P}_{11,i}(1)}, \\
\hat{p}_{2,i} &= \frac{1 - \hat{P}_{00,i}(2)}{2 - \hat{P}_{00,i}(2) - \hat{P}_{11,i}(2)}.
\end{aligned} \tag{5.12}$$

This method uses substantially more data than the single path method, as the sample size for the multipath method is s times that of the single path method.

Large values of the statistic $G^2_{\max}$ defined in equation (3.6) will indicate that a change is present in at least one of the sequences. Unfortunately, the multipath dependent LRT is only able to detect that a change has occurred in at least one of the sequences. This method cannot estimate the location(s) of the change point(s) nor identify which sequence(s) has actually changed.

Another limitation of this method is the lack of a known asymptotic distribution for the test statistic $G^2_{\max}$. A bootstrap procedure similar to the one described in Section 3.3.1 can provide approximate p-values. The run time of the single sequence bootstrap algorithm for modest sample sizes ($n \geq 250$) is quite long. The additional information used in a multiple sequence bootstrap procedure with no restrictions on the locations of the change points may

cause an unreasonable run time to generate a p-value.

Future work on the multipath dependent LRT includes finding an efficient bootstrap procedure or appropriate assumptions on the locations of the change points $\boldsymbol{\tau}$ to reduce the run time of the algorithm. The ideal result would be to identify the asymptotic distribution of the multipath $G^2_{\max}$ statistic.

5.3 Extensions to Sequences of Multinomial Trials

The tests described in this dissertation to detect and estimate change points in dependent sequences of Bernoulli random variables lay the framework for generalization to multinomial sequences with $K+1$ categories. Suppose that the sequence $y \sim \text{Multinomial}(1, \boldsymbol{p})$ with transition matrix $\boldsymbol{P}$. Here:

$$\boldsymbol{p} = \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_K \end{pmatrix} \quad \text{and} \quad \boldsymbol{P} = \begin{pmatrix} P_{00} & P_{01} & \cdots & P_{0K} \\ P_{10} & P_{11} & \cdots & P_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ P_{K0} & P_{K1} & \cdots & P_{KK} \end{pmatrix}.$$

The hypotheses of interest are:

H_0 : $x_t \sim \text{Multinomial}(1, \boldsymbol{p})$ with transition probabilities P_{uv} for all times t,

H_a : There exists τ, $1 < \tau < n$, such that $x_t \sim \text{Multinomial}(1, \boldsymbol{p}(1))$ for all $1 < t \leq \tau$ and $x_t \sim \text{Multinomial}(1, \boldsymbol{p}(2))$ for all $\tau < t \leq n$ where $\boldsymbol{p}(1) \neq \boldsymbol{p}(2)$ and the events after the change are independent of the events prior to the change.

Although the theoretical results for multinomial sequences are out of the scope of this dissertation, future work includes exploration of the asymptotic distributions of DCUSUM and Dependent LRT in the multinomial case. For convenience, these are briefly discussed in

Subsections 5.3.1 and 5.3.2.

5.3.1 Multinomial DCUSUM Test

The DCUSUM statistic is easily generalized to the multinomial case. Let $y \sim \text{Multinomial}(1, \boldsymbol{p})$ with transition matrix $\boldsymbol{P}$ and define $\boldsymbol{x}_t$ to be the observed value of y at time t. That is, $\boldsymbol{x}_t = (x_{0,t}, x_{1,t}, \ldots, x_{K,t})'$. The weighted sum for category k at time t is:

$$S_{k,t} = \sum_{j=1}^{t} x_{k,j} - \frac{t}{n}\sum_{j=1}^{n} x_{k,j} = \sum_{j=1}^{n} a_{k,j}x_{k,j}, \quad \text{where} \quad a_{k,j} = \begin{cases} 1 - \frac{t}{n} & \text{if } 1 \le j \le t, \\ -\frac{t}{n} & \text{if } t+1 \le j \le n. \end{cases}$$

The DCUSUM statistic for category k at time t is defined as:

$$\text{DCUSUM}_{k,t} = S_{k,t}/\sqrt{n}. \tag{5.13}$$

Notice that for each category, the $\text{DCUSUM}_{k,t}$ statistic is identical to the statistic given in equation (2.2) for the Bernoulli case. Because of this, it is believed that the multinomial DCUSUM statistic will converge to the same asymptotic distribution as in the independent multinomial case as stated by Robbins et al. [34]. This result is restated below, but the proof is reserved for future work.

Conjecture Suppose $y = \{\boldsymbol{x}_t\}_{t=1}^{n}$ where $\boldsymbol{x}_t \sim \text{Multinomial}(1, \boldsymbol{p})$ is an m-dependent sequence of multinomial random variables with $K+1$ categories and transition matrix defined by $\boldsymbol{P}$. Define $\sigma_k^2 = \text{Var}(x_{k,t})$ for each $k = 0, 1, \ldots, K$, then:

$$\max_{l \le t/n \le h} \left\{ \sum_{k=0}^{K} \sigma_k^{-2} \text{DCUSUM}_{k,t}^2 \Big/ \left(\frac{t}{n}\left(1 - \frac{t}{n}\right)\right) \right\} \xrightarrow{D} \sup_{l \le \eta \le h} \frac{B^{(K+1)}(t)}{\eta(1-\eta)},$$

where $B^{(K+1)}(t)$ is the sum of $K+1$ independent Brownian bridge processes.

This conjecture is the generalization of Theorem 2.3.1 and if proved true, p-value approximations for the test statistic may be found using equation (1.13) with $d = K+1$.

5.3.2 Multinomial Dependent LRT

Similar to the DCUSUM statistc, the log likelihood statistic $G^2_{\max}$ is generalizable to multinomial sequences. Let $y \sim \text{Multinomial}(1, \boldsymbol{p})$ with transition matrix $\boldsymbol{P}$ and define $\boldsymbol{x}_t$ to be the observed value of y at time t. Define n^t_{uv} as a generalization of the counts defined in equation (3.1), $n^{t^*}_{uv} = n^n_{uv} - n^t_{uv}$, and $x^*_{u,v,t,t+1} = \begin{cases} 1 & \text{if } x_t = u \text{ and } x_{t+1} = v, \\ 0 & \text{otherwise.} \end{cases}$

The modified likelihood function for a fixed time t under the null hypothesis is:

$$L^{**}_{H_0} = \prod_{u=0}^{K} \left(\left(\prod_{v=0}^{K-1} P_{uv}^{n^n_{uv} - x^*_{u,v,t,t+1}} \right) \left(1 - \sum_{v=0}^{K-1} P_{uv} \right)^{n^n_{uK} - x^*_{u,K,t,t+1}} \right),$$

while the modified likelihood function for a fixed time t under the alternative hypothesis is:

$$\begin{aligned} L^*_{H_a} = &\prod_{u=0}^{K} \left(\left(\prod_{v=0}^{K-1} P(1)_{uv}^{n^t_{uv}} \right) \left(1 - \sum_{v=0}^{K-1} P(1)_{uv} \right)^{n^t_{uK}} \right) \\ &\times \prod_{u=0}^{K} \left(\left(\prod_{v=0}^{K-1} P(2)_{uv}^{n^{t*}_{uv} - x^*_{u,v,t,t+1}} \right) \left(1 - \sum_{v=0}^{K-1} P(2)_{uv} \right)^{n^{t^*}_{uK} - x^*_{u,K,t,t+1}} \right). \end{aligned}$$

Optimizing the likelihood functions for each of the parameters yields a system of equations to calculate the MLEs for each of the k categories:

$$\begin{aligned} \hat{P}_{uk} &= \left(1 - \sum_{\substack{v=0 \\ v \neq k}}^{K-1} \hat{P}_{uv} \right) \left(\frac{n^n_{uv} - x^*_{u,v,t,t+1}}{n^n_{uv} - x^*_{u,v,t,t+1} + n^n_{uK} - x^*_{u,K,t,t+1}} \right) \quad \text{for } 1 \leq k < K, \\ \hat{P}_{uK} &= 1 - \sum_{v=0}^{K-1} \hat{P}_{uv}, \\ \hat{P}(1)_{uk} &= \left(1 - \sum_{\substack{v=0 \\ v \neq k}}^{K-1} \hat{P}(1)_{uv} \right) \left(\frac{n^t_{uv}}{n^t_{uv} + n^t_{uK}} \right) \qquad \text{for } 1 \leq k < K, \\ \hat{P}(1)_{uK} &= 1 - \sum_{v=0}^{K-1} \hat{P}(1)_{uv}, \end{aligned}$$

$$\hat{P}(2)_{uk} = \left(1 - \sum_{\substack{v=0 \\ v \neq k}}^{K-1} \hat{P}(2)_{uv}\right)\left(\frac{n_{uv}^{t^*} - x_{u,v,t,t+1}^*}{n_{uv}^{t^*} - x_{u,v,t,t+1}^* + n_{uK}^{t^*} - x_{u,K,t,t+1}^*}\right) \text{ for } 1 \leq k < K,$$

$$\hat{P}(2)_{uK} = 1 - \sum_{v=0}^{K-1} \hat{P}(2)_{uv}. \tag{5.14}$$

Closed form solutions of the MLEs may be found by solving the system (5.14). Once the MLEs are obtained, the G_t^2 statistic may be computed for each of the permissible times $nh \leq t \leq nl$ and the presence of large values of $G_{\max}^2 = \max_{nh \leq t \leq nl} G_t^2$ is evidence that a change has occurred in the sequence. The estimated location of the change point is $\hat{\tau} = \arg\max_{nh \leq t \leq nl} G_t^2$.

Future work includes obtaining closed form solutions to the system (5.14), exploration of the asymptotic distribution of G_t^2, implementation of a bootstrap procedure for p-values, and obtaining the asymptotic distribution of $G_{\max}^2$.

5.4 Application to Clustered Time Series Models

The initial motivation of the results in this dissertation was to provide a method of statistical inference to detect and estimate a change over time in the structure of a clustering scheme for clustered time series models. This section provides a review of time series fitting techniques as well as methods to cluster the fitted models. The theoretical results in Chapters 2 and 3 can then be applied to the clustering output to detect and estimate changes in the structure over time as described in Section 5.4.4.

5.4.1 Time Series Model Fitting

Define y to be a time series of length n, then y can be fit in a variety of ways depending on the assumptions of the model. The two procedures discussed below are the joinpoint regression model proposed by Kim et al. [19] and the multiple regression model.

If the only explanatory variable is a single time variable, the *joinpoint regression model* is preferred due to its change point detection and estimation capability. The model is defined in

Kim et al. [19] and is restated below. Consider the observations, $(t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)$, where $t_1 < t_2 < \cdots < t_n$, then:

$$\mu_y = \beta_0 + \beta_1 t + \delta_1 (t - \tau_1)^+ + \cdots + \delta_c (t - \tau_c)^+, \tag{5.15}$$

where $\tau_1, \ldots \tau_c$ are the unknown locations of c joinpoints, and the function $(x)^+$ is defined as:

$$(x)^+ = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

A permutation test or BIC value is used to determine the number of joinpoints by sequentially testing the hypotheses:

$$H_0 : \text{ there are } c_0 \text{ joinpoints,}$$

versus

$$H_a : \text{ there are } c_1 \text{ joinpoints.}$$

If the null hypothesis is rejected, then c_0 is increased by 1, otherwise c_1 is decreased by 1. This process is repeated until $c_0 = c_1 := c$. Initial values are generally chosen as $c_0 = 0$ and $c_1 = 5$. Once the number of joinpoints c is decided, the regression model (5.15) may be fit.

When the data consists of multiple predictors dependent on a time variable, the joinpoint model is not appropriate. Instead, a multiple regression is used to fit the model:

$$\mu_y = \boldsymbol{X}'\boldsymbol{\beta}, \tag{5.16}$$

where $\boldsymbol{X}$ is the design matrix and $\boldsymbol{\beta}$ is a vector of coefficients.

5.4.2 Clustering Methods

In order to apply a clustering scheme, it is necessary to have a reasonable number of time series. Let $\{y_i\}_{i=1}^{s}$ be a set of s time series. The clustering methods discussed below will use the fitted models from Section 5.4.1 to group the time series into $K+1$ clusters.

k-means Clustering

The k-means procedure is a non-parametric method used to group s independent N-dimensional observations into k groups. For consistency, define $k = K+1$. Note here that $s >> K+1$.

The process begins by arbitrarily assigning $K+1$ of the s points or vectors as cluster centers. Each additional observation is then assigned to the group whose mean is the smallest Euclidean distance away. The center of that cluster is updated to reflect the mean of all observations in that group. This continues until all points have been assigned to some cluster. The distance from each observation to each cluster center is then compared and objects are rearranged corresponding to that minimum distance. If there are ties in distance, the object is arbitrarily assigned to the cluster with smaller index. The process repeats until there is no change in cluster membership, or a certain number of iterations is reached.

Note that the number of clusters $K+1$ is assumed to be known. In practice, that is almost never the case. The Bayesian Information Criterion (BIC) will be used to determine the number of clusters, and is discussed in Section 5.4.3.

The term k-means generally refers to the work by MacQueen [23], even though his algorithm is not used in practice. The algorithm given by Hartigan and Wong [11] is more efficient than MacQueen's algorithm and has been implemented in several statistical packages. For the purposes of this paper, k-means will refer to Hartigan and Wong's algorithm.

This algorithm may be applied to both of the fitted time series models (5.15) and (5.16) in two ways. The first method is to use the vectors of coefficients $\boldsymbol{\beta}_i$ as cluster centers. This method will tend to group those observations with similar slopes together. The other method is to use the vectors of fitted values $\hat{\boldsymbol{y}}_i$ as cluster centers. This will group the time

series that have similar predicted values together.

Clustering of regression models (CORM)

The CORM method was developed in 2006 by Qin and Self [32]. This model based method was developed to cluster massive sequences of gene data over time, and is capable of dealing with multiple samples each at multiple time points.

Define $\boldsymbol{u}_i = (u_{i,0}, u_{i,1}, \ldots, u_{i,K})$ to be the cluster membership vector corresponding to time series y_i where exactly one $u_{i,k} = 1$ when y_i is an element of cluster k and the rest of the u's are 0. As described in Dempster et al. [6], the cluster membership probabilities $\boldsymbol{\pi}_i$ corresponding to the membership values $u_{i,k}$ are treated as missing data and the EM algorithm is used to estimate the values. For the purposes of this dissertation, data is restricted to single samples at multiple time points, which is described by Qin and Self [32] as longitudinal data with no replication (LNR).

The clustering algorithm assumes the following linear mixed effects model for LNR data. Let $\boldsymbol{y}_i$ be a vector of observed values, $\boldsymbol{X}_i$ the corresponding design matrix, ϵ_i the associated error term, $\boldsymbol{u}_i$ the cluster membership vector, and $\boldsymbol{\beta}_k$ the regression coefficients for cluster k. Assuming that there are $K+1$ clusters created by the objects that share the same values of the regression coefficients, the CORM model is:

$$
\begin{aligned}
&\boldsymbol{y}_i \mid (u_i = k, \boldsymbol{X}_i) = \boldsymbol{X}_i'\boldsymbol{\beta}_k + \epsilon_i, \\
&\epsilon_i \mid (u_i = k) \sim \text{MVN}(0, \boldsymbol{V}_i(\xi_k)), \\
&\boldsymbol{V}_i(\xi_k) = \boldsymbol{Z}_i\boldsymbol{D}_k\boldsymbol{Z}_i' + \sigma_k^2 I, \\
&u_i \sim \text{Multinomial}(\boldsymbol{\pi}_i).
\end{aligned}
$$

The variance $\boldsymbol{V}_i(\xi_k)$ can be thought of as the sum of the random effects and the measurement error. Notice that the model is capable of dealing with a different error term for each individual time series.

The initial estimates for $\boldsymbol{\beta}_k$ and $\boldsymbol{\pi}_i$ are found through random cluster assignments or a k-means procedure, similar to the method described in Section 5.4.2. The initial estimates should be found more than once to protect against focusing on a local maximum instead of a global maximum. After the initial estimates are made, the EM algorithm runs to fit more precise values of $\boldsymbol{\pi}_i$. The algorithm terminates once the increase of the log likelihood function from one iteration to the next is less than 0.01. The output considered are the membership probabilities organized in a membership matrix. The rows are viewed as individual multinomial random variables corresponding to each time series object.

In practice, the CORM method has two major limitations. First, the code available in the statistical package `R` assumes that the design matrices are the same for all s time series. This restricts the possibilities of models to fit substantially. The second limitation is the lack of a procedure to determine the number of clusters. This issue is addressed by using a BIC method discussed in Section 5.4.3.

When fitting the multiple regression time series model (5.16), the common design matrix does not effect the fitting process. On the other hand, when fitting the joinpoint time series model (5.15), some of the time series y_i may have a different number of joinpoints or different joinpoint locations. This may lead to different design matrices for each of the s time series. To work around this issue, a piecewise linear spline is used as the common design matrix. An example of the model with five basis functions is:

$$\mu_y = \beta_0 + \beta_1 t + \delta_1 \left(t - \frac{n}{6}\right)^+ + \delta_2 \left(t - \frac{2n}{6}\right)^+ + \cdots + \delta_5 \left(t - \frac{5n}{6}\right)^+ . \tag{5.17}$$

This model will be able to roughly estimate the unique joinpoint locations and slopes while providing a common design matrix $\boldsymbol{X}$ for all of the s time series.

5.4.3 Choosing a proper number of clusters using Bayesian information criterion (BIC)

The k-means and CORM clustering algorithms assume that the number of clusters for a given data set is known. In practice, that is very uncommon. There are several methods to determine an optimal number of clusters and here, the focus is on the Bayesian information criterion (BIC).

The BIC was first proposed by Schwarz [36] and has been widely accepted as one of the most useful tools to determine the optimal number of parameters in an arbitrary model. The criterion was developed from a Bayesian framework and made use of the posterior probabilities of a set of potential models.

The following review of the derivation of BIC and more detail about its uses is available in Knoishi and Kitagawa [35]. Suppose $C_1, C_2, \ldots, C_W$ are the potential clustering schemes. Each model C_k can be characterized by the distribution $f_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k)$ where $\boldsymbol{\theta}_k$ is a vector of potential cluster centers. Let $\pi(\boldsymbol{\theta}_k)$ represent the prior distribution of the parameter vector $\boldsymbol{\theta}_k$, then the marginal probability of $\boldsymbol{y}_s = \{y_1, \ldots, y_s\}$ is given as:

$$p_k(\boldsymbol{y}_s) = \int f_k(\boldsymbol{y} \mid \boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k. \tag{5.18}$$

The basic definition of Bayes theorem states:

$$P(C_k \mid \boldsymbol{y}_s) = \frac{p_k(\boldsymbol{y}_s)P(C_k)}{\sum_{j=1}^{W} p_j(\boldsymbol{y}_s)P(C_j)}. \tag{5.19}$$

where $P(C_j)$ denotes the prior probability that clustering scheme C_j is selected. The goal is to maximize (5.19), and if prior probabilities are assumed equal, this is equivalent to maximizing (5.18). Because $p_k(\boldsymbol{y}_s)$ can be thought of as a likelihood, it is intuitive from a

statistical viewpoint to consider the following transformation:

$$\begin{aligned} -2\log p_k(\boldsymbol{y}_s) &= -2\log\left\{\int f_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k)\pi(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k\right\} \\ &\approx -2\log f_k(\boldsymbol{y}_s \mid \hat{\boldsymbol{\theta}}_k) + c_k \log s. \end{aligned} \tag{5.20}$$

Here, $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator of $\boldsymbol{\theta}_k$ and c_k is the number of parameters in the vector $\boldsymbol{\theta}_k$.

The value of equation (5.20) is the BIC value for the k^{th} clustering scheme. That is:

$$\mathrm{BIC}_k = -2\log f_k(\boldsymbol{y}_s \mid \hat{\boldsymbol{\theta}}_k) + c_k \log s.$$

It is made up of two components, the logarithm of the maximum likelihood and a penalty function to limit the size of the model chosen. From this definition, the optimal number of clusters $\hat{K}$ is defined as:

$$\hat{K} = \underset{1\leq k\leq W}{\arg\min}\, \mathrm{BIC}_k. \tag{5.21}$$

5.4.4 Detection and Estimation of a Change Point in Clustered Time Series Models

Consider s time series models $\{y_i\}_{i=1}^s$ of length n and define $c_{\hat{K}}$ to be the maximum number of parameters in the design matrix, where $\hat{K}$ is determined by the minimum BIC value discussed in Section 5.4.3. Define $t_0 > c_{\hat{K}}$ and for each time $t = t_0, t_0+1, \ldots, n$, consider the time series $\{y_{i,t}\}_{i=1}^s$ truncated at time t. (The choice of t_0 is made to prevent from overfitting the time series models.) The truncated time series are fit using either a joinpoint or multiple regression model and then clustered using k-means clustering or CORM.

If k-means clustering is used, the cluster membership values are already defined as integer values, while if CORM is used, the cluster membership probabilities $\boldsymbol{\pi}_i$ are converted into

integer values by assigning membership to the cluster with largest probability $\pi_{i,k}$. In either case, the resulting cluster membership values at each time t can be viewed as the elements of a Bernoulli or multinomial sequence of random variables. The change point detection and estimation techniques proposed in this dissertation may then be applied to the sequences of random variables to determine if and when a change occurred in the clustering scheme.

Under the null hypothesis of no change in the cluster structure, it is more likely that each time series y_i will stay in the same cluster from time t to time $t+1$. This was the main motivation for assuming a one step Markov dependence in the sequences of random variables for the change point detection and estimation procedures discussed in Chapters 2 and 3.

In order for this method to be useful in practice, the multinomial and multipath techniques in Sections 5.2 and 5.3 must be explored further. Future work includes coding a program to link the clustering output of the truncated time series for each time t to the existing detection and estimation programs, as well as exploration of the robustness of using estimated cluster membership values instead of the population parameter values in change point detection and estimation.

5.5 Other Applications

There are several possible real world applications for both single and multi path techniques discussed throughout this dissertation. The possibilities are vast; however, only a few are mentioned here to motivate the use of these dependent methods.

The single path applications detect and estimate a change in only one time series y. Consider the random variables x_t to indicate whether or not a product produced by a machine is defective at equally spaced time intervals. For quality control purposes, the detection of a change point may indicate that the machine needs repair. Another application is to consider the variables x_t to be indicators of annual party majority in American congress (where, say 1 represents democrats, while 0 represents republicans). A statistically significant change

would indicate a political shift, and in the case of other countries with more than two leading parties, the multinomial procedures may be implemented.

Multi path applications are capable of utilizing multiple time series $\{y_i\}_{i=1}^{s}$ simultaneously. One use would be to cluster the daily, monthly, or annual closing prices of the S&P 500 surrounding the recession of the late 2000s to better understand which companies were effected and detect similarities and differences in industry movement. In general, the multi path procedures may be used to better understand financial markets from real estate to automobile sales, and any other financial data of interest.

5.6 Concluding Remarks

Motivated by the desire to explore clusters of time series models, the single path DCUSUM and dependent LRT procedures provide change point detection and estimation techniques for sequences of dependent Bernoulli random variables. The known asymptotic distribution of the DCUSUM statistic and p-value approximation methods provide a mechanism for generating p-values, while a bootstrap procedure is necessary for hypothesis testing using the dependent LRT method. The extensions to multipath and/or multinomial DCUSUM and dependent LRT procedures, as well as the maximal change count statistic Δ_t discussed in this chapter, provide a basis for future research on the topic of change point detection and estimation in dependent sequences of random variables.

Bibliography

[1] Roger L. Berger and Dennis D. Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):pp. 1012–1016, 1994.

[2] Patrick Billingsley. *Statistical Inference for Markov Processes*. The University of Chicago, 1961.

[3] G. Casella and R.L. Berger. *Statistical inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[4] Yun-Shiow Chen. The change-point problem in a multinomial sequence and its application, 1988. Copyright - Copyright UMI - Dissertations Publishing 1988; Last updated - 2014-01-22; First page - n/a; M3: Ph.D.

[5] K.L. Chung. *A Course in Probability Theory*. Academic Press, 2001.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38, 1977.

[7] Jay L. Devore. A note on the estimation of parameter in a bernoulli model with dependence. *The Annals of Statistics*, 4:990–992, 1976.

[8] R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[9] Paul I. Feder. The log likelihood ratio in segmented regression. *The Annals of Statistics*, 3(1):pp. 84–97, 1975.

[10] Aaron L. Halpern. Minimally selected p and other tests for a single abrupt changepoint in a binary sequence. *Biometrics*, 55(4):pp. 1044–1050, 1999.

[11] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):pp. 100–108, 1979.

[12] D. V. Hinkley. Inference about the intersection in two-phase regression. *Biometrika*, 56(3):pp. 495–504, 1969.

[13] David V. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):pp. 1–17, 1970.

[14] David V. Hinkley and Elizabeth A. Hinkley. Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):pp. 477–488, 1970.

[15] Wassily Hoeffding and Herbert Robbins. The central limit theorem for dependent random variables. *Duke Math. J.*, 15(3):773–780, 09 1948.

[16] Joel L. Horowitz. Chapter 52 - the bootstrap. volume 5 of *Handbook of Econometrics*, pages 3159 – 3228. Elsevier, 2001.

[17] Lajos Horvth and Monika Serbinowska. Testing for changes in multinomial observations: The lindisfarne scribes problem. *Scandinavian Journal of Statistics*, 22(3):pp. 371–384, 1995.

[18] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.

[19] Hyune-Ju Kim, Michael P. Fay, Eric J. Feuer, and Douglas N. Midthune. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, 19(3):335–351, 2000.

[20] J. Krauth. Cchange-point in bernoulli trials with dependence. *Between Data Science and Everyday Web Practice*, pages 261–269, 2003.

[21] J. Krauth. Test for a change point in bernoulli trials with dependence. *Innovations in ClaInnovations, Data Science and Information Systems*, pages 154–164, 2005.

[22] Ian B. MacNeill. Tests for change of parameter at unknown times and distributions of some related functionals on brownian motion. *The Annals of Statistics*, 2(5):pp. 950–962, 1974.

[23] J. MacQueen. Some methods for classification and analysis of multivariate observations, 1967.

[24] E. Mammen. *When does bootstrap work? : asymptotic results and simulations*. New York : Springer-Verlag, 1992.

[25] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 03 1947.

[26] Rupert Miller and David Siegmund. Maximally selected chi square statistics. *Biometrics*, 38(4):pp. 1011–1016, 1982.

[27] Steven Orey. A central limit theorem for m -dependent random variables. *Duke Math. J.*, 25(4):543–546, 12 1958.

[28] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):pp. 100–115, 1954.

[29] E. S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):pp. 523–527, 1955.

[30] Lusa Pereira. On the extremal behavior of a nonstationary normal random field. *Journal of Statistical Planning and Inference*, 140(11):3567 – 3576, 2010.

[31] A. N. Pettitt. A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika*, 67(1):pp. 79–84, 1980.

[32] Li-Xuan Qin and Steven G. Self. The clustering of regression models method with applications in gene expression data. *Biometrics*, 62(2):pp. 526–533, 2006.

[33] Alex Riba and Josep Ginebra. Change-point estimation in a multinomial sequence and homogeneity of literary style. *Journal of Applied Statistics*, 32(1):61–74, 2005.

[34] Michael W. Robbins, Robert B. Lund, Colin M. Gallagher, and QiQi Lu. Change-points in the north atlantic tropical cyclone record. *Journal of the American Statistical Association*, 106(493):89–99, 2011.

[35] Genshiro Kitagawa Sadanori Knoishi. Bayesian information criteria. In *Information Criteria and Statistical Modeling*, Springer Series in Statistics, pages 211–213. Springer New York, 2008.

[36] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):pp. 461–464, 1978.

[37] Y. H. Wang. On the number of successes in independent trials. *Statistica Sinica*, 3:295 – 312, 1993.

[38] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):pp. 60–62, 1938.

[39] Douglas A. Wolfe and Yun-Shiow Chen. The changepoint problem in a multinomial sequence. *Communications in Statistics - Simulation and Computation*, 19(2):603–618, 1990.

[40] K. J. Worsley. An improved bonferroni inequality and applications. *Biometrika*, 69:297–302, 1982.

[41] K. J. Worsley. The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika*, 70(2):pp. 455–464, 1983.

BIOGRAPHICAL DATA

NAME OF AUTHOR: Benjamin D. Cortese

PLACE OF BIRTH: Binghamton, NY

DATE OF BIRTH: November 2, 1987

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

The College at Brockport: State University of New York - Brockport, NY

Syracuse University - Syracuse, NY

DEGREES AWARDED:

B.S., The College at Brockport: State University of New York - Brockport, 2009

M.A., Syracuse University - Syracuse, 2011