12-2006

# Change-Point Tests for Precipitation Data

Michael Robbins
*Clemson University*, mrobbin@clemson.edu

# THE LIKELIHOOD APPROACH IN PRECIPITATION CHANGE-POINT TESTING

---

A Thesis
Presented to
the Graduate School of
Clemson University

---

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Sciences

---

by
Michael Robbins
December 2006

---

Accepted by :
Dr. Colin M. Gallagher, Committee Chair
Dr. Robert B. Lund
Dr. K.B. Kulasekera

# ABSTRACT

A new method is required for change-point testing of precipitation data that is capable of applying valid precipitation models. First, stochastic precipation models are researched and classified. Typically, the occurrence of rain is modeled using a two-state, first-order Markov chain, and the intensity of rain is modeled using a two-parameter gamma distribution. Using the likelihood ratio test statistic, methods are devoloped for testing for fixed and unknown change-points. These methods are developed for various models, including the MC/gamma model and simplified versions. The distribution of the LRT is unknown, however its asymptotic distribution is known for both the fixed and unknown change-point tests. First, the asymptotic converegence rates are analyzed using simulation, and then the power of the test is also analyzed using simulation. Finally the test is applied to real data, and the results are analyzed.

# TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF TABLES

# Chapter 1

# Introduction

In any set of data that can viewed as a time series, be it climatological or otherwise, one of the more important methods of analyses is to test for change-points, or a lack of homogeneity. For climatological data, a change-point may be instigated by a climate shift (such as global warming) or by a simple change in the method of data collection (such as the implementation of a new weather gauge). The detection of these change-points can not only provide vast insight into past weather behavior, but can also improve forecast accuracy. At its most propitious, a test could determine whether or not a perceived change-point in climate behavior is instigated by something as mundane as a weather gauge or something as ubiquitously encompassing as a drastic climate shift.

Change point testing usually involves either the verification of a change-point at fixed time in the data or the detection of a change-point at an unknown time over a wide range of data. Testing for non-fixed change points is much more complicated, and the exact statistical quantiles that are used usually have to be numerically estimated.

In time series analysis, one of the most common and useful assumptions is that the data come from, to at least some extent, a normal distribution. Under the assumption

of normality, the problem of change-point testing is so well studied that, regardless of the application and methodology, an accurate change point test is very easily obtained. For instance, even though temperature series show a very high correlation and display a seasonal trend that cannot be disregarded, simple time series adjustments can be made such that the only random element in the data is due to white noise.

If there is one climatological quantity that is stochastically modeled, in which normality simply cannot be assumed, it is precipitation data. In monthly and even annual precipitation data, normality is quite a stretch. In daily rainfall data, in which observations above zero may occur in less than 10% of recorded days, the assumption of normality is simply impossible. This is a simple fact of stochasitics that is nevertheless ignored in one of the most predominate methods of homogeneity tests for precipitation.

In this test, Alexandersson (1986), the amount of rain on a given day is notated as $Y_i$, and the ratio $q_i$ is defined:

$$q_i = \frac{Y_i}{X_i},$$

where $X_i$ is the corresponding observed rainfall at a nearby station. The $\{q_i\}$ are then standardized to

$$Z_i = \frac{q_i - \bar{q}}{\hat{\sigma}},$$

where $\bar{q}$ is the sample mean of the ratios and $\hat{\sigma}$ is the standard deviation of the ratios. Each $Z_i$ is assumed to come from a standard normal distribution.

This method of testing is so contrived that, from a statistical standpoint, the validation of its assumptions is, if not impossible, extremely difficult, and the significance of its results are, if not meaningless, practically incomprehensible. Even in this testing method, each of the observations represent annual rainfall. A similar test for daily rainfall would be vastly

more ridiculous. The aim of this study is to present change-point testing methods under the more valid models of precipitation in which normality is not assumed.

First, we explore the preferred statistical rainfall models and discover that popularly, daily precipitation is broken in two parts which are separately estimated: the occurrenence of rain, and the amount of rain that is observed on rainy days. We begin our study by using a simplified version of the more complex models: one (refered to as the $MExp$ model) that assumes that rainfall on one day is independent of all other days, and that rain occurs with a probability, $p$, and that the amount of rain on rainy days is sampled from an exponential distribution. First, we present the methodology for testing for a fixed changed point at time $c^*$. The statistic used is the likelihood ratio test statistic, denoted $\Lambda_{c^*}$. Fortunately, $-2\log(\Lambda_{c^*})$ has an asymptotic distribution, which can provide quantiles for the test. When testing for an unknown change-point, which can occur at any admissible time $c$, the test statistic that is used is $\Lambda = \min_{l \le c \le h} \{\Lambda_c\}$. An asymptotic distribution of $-2\log(\Lambda)$ can be found, which again provides quantiles for the test. The power of the test is examined, and similar methods under more complicated precipitation models are presented. Finally, the methods are applied to a real data set, and the results are analyzed.

# Chapter 2

# Stochastic Models for Precipitation

Most commonly, annual and monthly precipitation is modeled using a gamma distribution, Wilks (2006), where the observations are assumed to be independent. For daily rainfall, in which a significant frequency of observations are 0, things get significantly more complex.

For a length-$n$ series of rainfall data, $\{X_i\}_{i=1}^{n}$, where $X_i$ represents the observed rainfall on day $i$ (may be 0), typically each $X_i$ is represented as the product of two random variables, $J_i$ and $Z_i$, such that

$$X_i = J_i Z_i,$$

where

$$J_i = \begin{cases} 1 & \text{if the } i^{th} \text{ day has rain,} \\ 0 & \text{if the } i^{th} \text{ day is dry.} \end{cases}$$

$Z_i$ is the amount of rainfall on day $i$ (given that day $i$ sees rain).

## 2.1 Models for $J_i$, the Occurrence of Rain

Typically, the sequence $\{J_i\}_{i=1}^n$ is thought to behave like a two-state, first-order Markov chain, where the chain is in state 0 at time $i$ if $J_i = 0$ and in state 1 at time $i$ if $J_i = 1$, Katz (1977). The probability matrix is

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

The first-order Markov chain assumption implies that

$$P\left[J_k = j | J_0, J_1, \ldots, J_{k-2}, J_{k-1} = i\right] = P\left[J_k = j | J_{k-1} = i\right].$$

Therefore

$$p_{00} = P\left[J_k = 0 | J_{k-1} = 0\right],$$

$$p_{01} = P\left[J_k = 1 | J_{k-1} = 0\right],$$

$$p_{10} = P\left[J_k = 0 | J_{k-1} = 1\right],$$

$$p_{11} = P\left[J_k = 1 | J_{k-1} = 1\right].$$

For instance, $p_{10}$ is the probaillity that it rains on any given day where it is known that it did not rain the previous day, etc. Because it is a two-state chain, we know

$$p_{01} = P\left[J_k = 1 | J_{k-1} = 0\right] = 1 - P\left[J_k = 0 | J_{k-1} = 0\right],$$

$$p_{01} = P\left[J_k = 1 | J_{k-1} = 1\right] = 1 - P\left[J_k = 0 | J_{k-1} = 1\right].$$

So if we let

$$p_0 = P\left[J_k = 1 | J_{k-1} = 0\right], \text{ and } p_1 = P\left[J_k = 1 | J_{k-1} = 1\right],$$

then the probability transition matrix can be rewritten as

$$\mathbf{P} = \begin{pmatrix} (1-p_0) & p_0 \\ (1-p_1) & p_1 \end{pmatrix}.$$

The stationary probabilities, $\pi_0$ and $\pi_1$, are defined as $\pi_i = \lim_{k\to\infty} P\left[J_k = i\right]$. For instance, $\pi_0$ can be thought of as the long-run proportion of days that are dry, and $\pi_1$ can be thought of as the long-run proportion of days that are wet. These probabilities are given by

$$\pi_0 = \frac{1-p_1}{1-p_1+p_0},\ \pi_1 = \frac{p_0}{1-p_1+p_0}.$$

However, stochastics is by definition uncertain, so no stochastic model is ever universally accepted. Hence, some climate researchers prefer to model precipitation with a two-state, second-order Markov chain, Stern & Coe (1984), where

$$P\left[J_k = j | J_0, J_1, , J_{k-2}, J_{k-1}\right] = P\left[J_k = j | J_{k-2}, J_{k-1}\right].$$

Not all climatologists believe that even the $2^{nd}$-order Markov chain can sufficiently model precipitation, so some researchers seek an $m^{th}$-order chain, Gregory & Jones (1992), such that

$$P\left[J_k = j | J_0, J_1, \ldots, J_{k-2}, J_{k-1}\right] = P\left[J_k = j | J_{k-m}, \ldots, J_{k-1}\right].$$

The simplest model for the occurence of precipitation is the one-parameter model where

$$P\left[J_k = 1 | J_0, J_1, , J_{k-2}, J_{k-1}\right] = P\left[J_k = 1\right] = p;$$

$$P\left[J_k = 1 | J_0, J_1, , J_{k-2}, J_{k-1}\right] = P\left[J_k = 0\right] = 1 - p,$$

such that $p$ is simply the probability of any given day being rainy. This model assumes that a rainfall series, $\{X_i\}_{i=1}^{n}$, is an indepedent and identically distributed set.

## 2.2 Models for $Z_i$, the Amount of Precipitation on Rainy Days

When it rains, the amount of rain, $Z_i$, is commonly considered to be independent of other days, Katz (1999). Different researchers propose different distributions for $Z_i$. Here are a few of the most common:

-The $Weibull(\gamma, \beta)$ distribution, Chapman (1997), with desity function

$$f(x|\gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-x^{\gamma}/\beta}, \quad 0 \leq x < \infty, \quad \gamma > 0, \quad \beta > 0.$$

-The $Lognormal(\mu, \sigma^2)$ distribution, Burgueno & Lana (2004), with density function

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \frac{e^{-(\log x - \mu)^2/(2\sigma^2)}}{x}, \quad 0 \leq x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

-The $Skew\ Normal(\alpha)$ distribution, Harmel (2000), with density function

$$f(x|\alpha) = \sqrt{\frac{2}{\pi}} e^{-x^2/2} \int_{-\infty}^{\alpha x} e^{-t^2/2}/\sqrt{2\pi} dt.$$

-The three-parameter $Kappa(\alpha, \beta, \theta)$ distribution, Mielke & Johnson (1973), with density function

$$f(x|\alpha,\beta,\theta) = \frac{\alpha\theta}{\beta}\left(\frac{x}{\beta}\right)^{\theta-1}\left[\alpha + \left(\frac{x}{\beta}\right)^{\alpha\theta}\right]^{-(\alpha+1)/\alpha}, \quad 0 \le x < \infty, \quad \alpha,\beta,\theta > 0.$$

-The $Mixed\ Exponential(\alpha, \beta_1, \beta_2)$ distribution, Wilks (1999), with density function

$$f(x|\alpha,\beta_1,\beta_2) = \frac{\alpha}{\beta_1}e^{-x/\beta_1} + \frac{1-\alpha}{\beta_2}e^{-x/\beta_2}, \quad 0 \le x < \infty, \quad 0 \le \alpha \le 1, \quad \beta_1,\beta_2 > 0.$$

The most commonly used rainfall distribution is $gamma(\alpha, \beta)$, Stern & Coe (1984), with density function

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta}, \quad 0 \le x < \infty, \quad \alpha,\beta > 0.$$

When the Markov chain structure of $\{J_i\}_{i=1}^{n}$ is ignorable, it is usually assumed that each $X_i$ is $iid$ with cumulative distribution function

$$F(x) = (1-p) + pG(x),$$

where $G(x)$ is the $cdf$ for a $gamma(\alpha, \beta)$ distribution, Katz (1999).

When the parameters show a strong seasonal structure, Fourier series are used to account for the seasonality, Woolhiser & Pegram (1979), Stern & Coe (1984).

Often, the appropriate distribution depends upon the location of the data collection. For instance, Wilks (1998) indicated that the mixed exponential distribution could provide good fits to non-zero daily precipitation data in New York State, USA, whereas the gamma distribution is shown to be much more appropriate in other locations.

## 2.3 Other Modeling Techniques

Sometimes, it is not assumed that $\{Z_i\}_{i=1}^n$ is an independent series. One can postulate that the distribution of the amount of precipitation on a wet day depends upon whether or not the previous day was wet or dry. This implies

$$P\left[Z_k \leq x | J_0, J_1, \ldots, J_{k-2}, J_{k-1} = i\right] = P\left[Z_k \leq x | J_{k-1} = i\right],$$

where the conditional distribution functions, $F_0(x)$ and $F_1(x)$, are defined such that

$$F_0(x) = P\left[Z_k \leq x | J_{k-1} = 0\right] \neq F_1(x) = P\left[Z_k \leq x | J_{k-1} = 1\right].$$

Frequently, $F_0(x)$ and $F_1(x)$ are thought to have separate gamma distributions, such that

$$\frac{d}{dx} F_i(x|\alpha_i, \beta_i) = f_i(x|\alpha_i, \beta_i) = \frac{1}{\Gamma(\alpha_i)\beta_i^{\alpha_i}} x^{\alpha_i - 1} e^{-x/\beta_i}$$

for $i = 1$ *or* 2, Katz (1977).

A highly abstract precipitation model is the Srikanthan & McMahaon (2001) model which extends the Markov chain concept to a multi-state model described by its transition probability matrix. The daily rainfalls are grouped into up to 7 classes of given magnitude ranges, and the probabilities are calculated for the transition from each class to any other. The lowest class gives the occurrences of dry days, the top class is modeled by a skewed normal distribution, and intermediate classes are modeled by a linear distribution.

When daily rainfall is modeled using the *cdf*, $F(x) = 1 - p + pG(x) = q + (1 - q)G(x)$ where $q = 1 - p$, $0 \leq p \leq 1$, and where $G$ is gamma, we can say that $F$ has a modified (or mixed) gamma distribution, Thom (1968). This enables us to think of the distribution of

rainfall over a period of days $(x_1, \ldots, x_n)$ as the sum of modified gammas, where

$$F\left(\sum_{i=1}^{n} x_i\right) = q \cdots q + (1 - q \cdots q) G\left(\sum_{i=1}^{n} x_i\right),$$

and $G(\sum x_i)$ has a near gamma distribution, Thom (1968).

In Wilks (1990), it is assumed that $X_i > 0$ for all $i$. This means that even for arid climates there is actually rain every day. This enables him to treat all precipitation observations as being independent and identically distributed from the same gamma distribution. However, observations of no rain are treated as being position observations that are less than a pre-specified censoring value, $C$. This value, $C$, is usually the minimum value that the rain sensor is able to detect. The magnitude of $C$ is frequently very small. These assumptions presume that, in arid climates, daily rainfall follows a distribution where a very large percentage of the density is less than some very small number. Even if this is a valid charateristic of a proper distribution, it seems hard to believe that the observations of larger rainfalls would follow that same density.

## 2.4 The $MExp$ Distribution

Most of the models display an unnecessary degree of complexity for the purposes of testing for change-points. In order to develop a method of change-point testing, the following model for precipitation, which is a simplification of some of the more complex models described above, is introduced.

The amount of rain on any given day follows a modified exponential distribution,

$MExp(p, \lambda)$, where

$$X \sim MExp(p, \lambda) \Rightarrow P(X \leq x) = (1 - p) + p(1 - e^{-\lambda x}), \quad x \geq 0.$$

It is assumed that the amount of rain on any day is independent of rain amounts on other days. This model ignores the Markov chain behavior of precipitation and assumes that it rains with probability, $p$. When it does rain, the amount of rain follows an exponential distribution (which is a gamma distribution with shape parameter 1).

Using the rain data that will be used in this paper, Figure 2.1 shows a histogram of the amount of observed rain on 2823 rainy days.

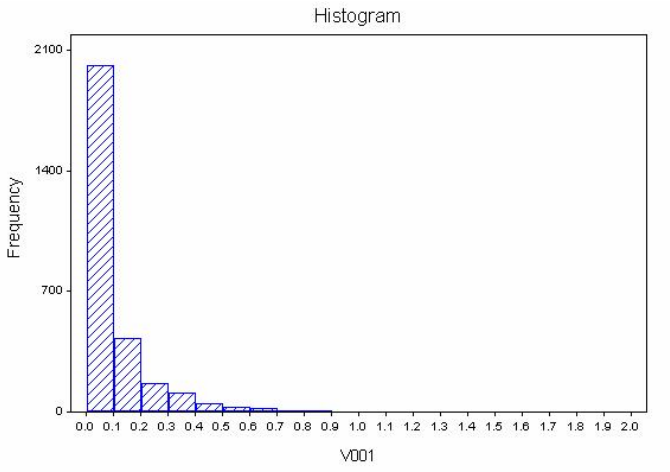


Figure 2.1: Histogram of Rainfall on Rainy Days

If these observations are assumed to come from an exponential distribution with mean $1/\lambda$, then we find

$$\hat{\lambda} = 10.15$$

2823 simulated observations of an exponential distribution with mean 1/10.15 give the histogram shown in Figure 2.2.

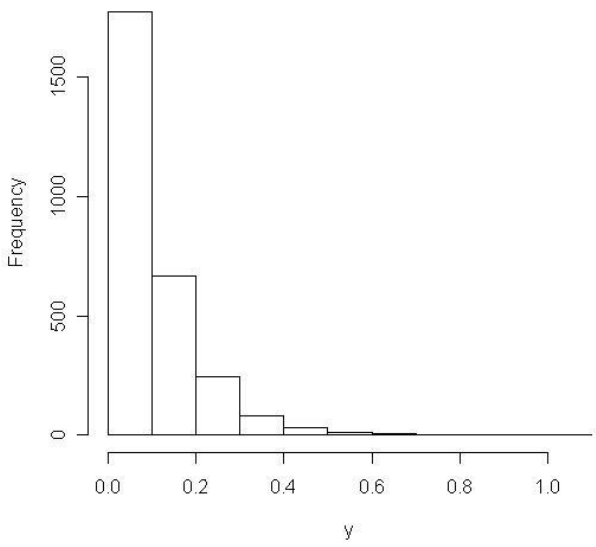


Figure 2.2: Histogram of Simulated Rainfall on Rainy Days (exp)

The actual max observation over the 2823 days or rain is 1.95; however, when rain is simulated using an exponential distribution, we find the maximum of the simulated observations is 1.06. Clearly the exponential distribution is a good fit, aside from the fact that it underestimates the large amounts of rain. Popularly, climatologists prefer to assume that rain observations have a gamma distribution, as opposed to an exponential one. So how much improvement is gained by modeling this rain with an gamma model? If a gamma distribution is assumed, we find

$$\hat{\alpha} = 0.803, \quad \hat{\beta} = 8.15.$$

The method of estimating parameters when a gamma distribution is assumed will be discussed later. When 2823 observations of a gamma distribution with the above parameters are simulated, the histogram in Figure 2.3 is produced.

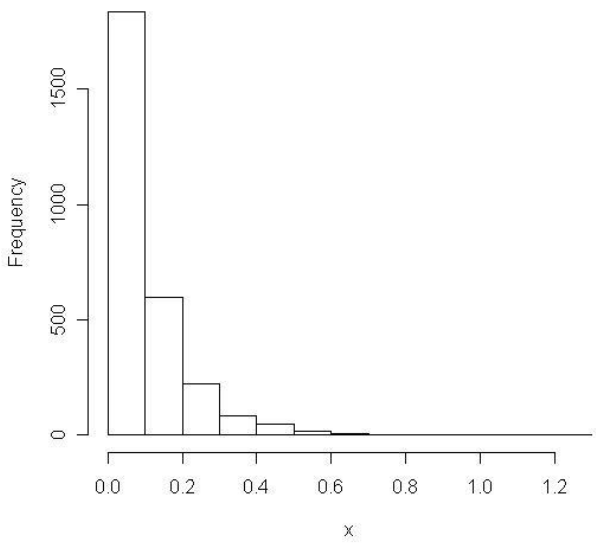


Figure 2.3: Histogram of Simulated Rainfall on Rainy Days(gamma)

The maximum of the simulated values is 1.29, which is still rather far from the observed maximum of 1.95. Hence, assuming a gamma distribution results in little improvement which, for the purposes of this paper, is not enough to make up for the plethora of complications that doing so adds.

# Chapter 3

# Precipitation Model Classification

If $\{X_i\}_{i=1}^n$ is a length-$n$ precipitation data series that, for each $i$, can be decomposed by letting $X_i = Z_i J_i$, then we define the following classification system for $\{X_i\}_{i=1}^n$:

## 3.1 Type A & B Classification

A precipitation model is of Type A if both $\{J_i\}_{i=1}^n$ and $\{Z_i\}_{i=1}^n$ are considered to be *iid* series, which most importantly implies that $\{X_i\}_n^{i=1}$ is *iid*. It is therefore imposed that for all $i$, $J_i \sim Bernoulli(p)$. Also, for all $i$, $Z_i$ has distribution function $G(z)$, defined for $0 < z < \infty$. Therefore, a Type A precipitation model assumes that $X_i$ has distribution function, $F(x|p,\theta) = (1-p) + pG(x|\theta)$ for each $i$. Inconveniently, $F(x)$ is a mixture distribution, with both discrete and continuous parts, which makes the definition of its density function conceptually difficult. The $MExp(p,\lambda)$ distribution that will be prevalent throughout this paper is a Type A model. The $MGam(p,\alpha,\beta)$ is a Type A model where $G(x)$ has a $Gamma(\alpha,\beta)$ distribution fuction. This model will be discussed in further detail later in this study.

Also, the model used by Wilks (1990), where it is assumed that daily rainfall amounts

all come from the same gamma distribution (and observations of 0 are considered to have positive values that are less than a small cut-off, C), is a Type A model where $p$ is restricted to 1. Likewise, models for monthly and annual precipitation data series, that are entirely non-zero, can also be thought of as Type A models with $p$ restricted to 1.

A model is said to be of Type B if $\{J_i\}_{i=1}^n$ is considered to behave as a two-state, first-order Markov chain and if $\{Z_i\}_{i=1}^n$ is $iid$ where $Z_i$ has distribution $G(z|\theta)$. This is a parametric model where the transition probabilities can be estimated. This is the most wide-spread precipitation model in climate research.

## 3.2 Further Classifications

The Type C classificiation holds the first-order, two-state Markov chain assumption for $\{J_i\}_{i=1}^n$ but drops the independence assumption in $\{Z_i\}_{i=1}^n$, typically by assuming that the distribution of $Z_i$ depends on $J_{i-1}$. The Type D classification models $\{J_i\}_{i=1}^n$ using a two-state Markov chain of order two or greater. The Type E classification models precipitation using a Markov chain with more than two states. The Srikanthan and McMahon model is a Type E model.

If the data shows a seasonal trend, the assumption of an identical distribution can be dropped in any of these classifications, and seasonality in the parameters can be accounted for.

This paper will only develop change-point methodology for Types A and B models, however Types C, D and E are all parametric models to which the methods in the paper could be adapted. However, a Type $\Omega$ precipitation model is one that applies non-parametric assumptions. For more information regarding change-point testing for non-parametric models, see Csorgo & Horvath (1997).

# Chapter 4

# The Likelihood Ratio

The most common test method used in parametric based change-point testing is the likelihood ratio test [Wilks (2006); Casella & Berger (2002); Csorgo & Horvath (1997)]. The LRT presents such an important methodology because it is a highly versatile test that presents a method of accepting or rejecting almost any null hypothesis against almost any alternative hypothesis. For instance, in the case of general change-point testing, let the series $X_1, X_2, \ldots, X_n$ be independent random variables with the respective distribution fucntions, $F(x;\theta_1), \ldots, F(x;\theta_n)$. We wish to test the null hypothesis

$$H_0 : \theta_1 = \ldots = \theta_n$$

against the alternative

$$H_A^* : \theta_1 = \ldots = \theta_{c^*} \neq \theta_{c^*+1} = \ldots = \theta_n$$

The null hypothesis above assumes that there is no change-point in the series, whereas the alternative assumes that there is a change-point at a known time $c^*$.

A more complicated version (and more useful in applications) is one that tests for a change-point at some unknown time $c$, with the alternative hypothesis becoming

$$H_A : \text{there is some integer } c,\ l \leq c \leq h,\ \text{such that}$$

$$\theta_1 = \ldots = \theta_c \neq \theta_{c+1} = \ldots = \theta_n.$$

The values $l, \ldots, h$ represent the range of possible times at which the change-point may have occured. Note, of course, that $1 \leq l \leq h \leq n$.

When testing for a fixed change-point, the likelihood ratio, $\Lambda_{c^*}$ rejects $H_0$ for small values, where

$$\begin{aligned} \Lambda_{c^*} &= \frac{\sup_\theta \prod_{i=1}^{n} f(X_i;\theta)}{\sup_{\theta_1} \prod_{i=1}^{c^*} f(X_i;\theta_1) \cdot \sup_{\theta_2} \prod_{i=c^*+1}^{n} f(X_i;\theta_2)} \\ &= \frac{\prod_{i=1}^{n} f(X_i;\hat{\theta})}{\prod_{i=1}^{c^*} f(X_i;\hat{\theta}_1) \cdot \prod_{i=c^*+1}^{n} f(X_i;\hat{\theta}_2)}. \end{aligned}$$

## 4.1 Type A Models and the LRT

If the series $X_1, X_2, \ldots, X_n$ represents daily rainfall amounts and if we assume a Type A precipitaion model, then for each $i$, $X_i$ has the distribution $F(x|p,\theta) = (1-p) + pG(x|\theta)$. Where the rain intensity is thought to be distributed $G(x|\theta)$ with $\theta$ being a vector of parameters. $F(x|p,\theta)$ is a mixture of discrete and continuous random variables and is

defined to have the following pdf

$$f(x;p,\lambda) = \begin{cases} p \cdot g(x|\theta) & \text{if } x > 0, \\ 1-p & \text{if } x = 0, \\ 0 & \text{else,} \end{cases}$$

where

$$g(x|\theta) = \frac{d}{dx} G(x|\theta).$$

As stated previously, let

$$J_i = \begin{cases} 1 & \text{if } X_i > 0, \\ 0 & \text{else.} \end{cases}$$

And now let

$$n_r = \sum_{i=1}^{n} J_i.$$

It is best to think of $n_r$ as the number of days in which there was rain (out of $n$ total days in the series), and now the maximum likelihood estimate for the parameter $p$ can be found:

$$\hat{p} = \frac{n_r}{n}.$$

Also, define

$$\mathbf{y} = \{Y_k\}_{k=1}^{n_r},$$

so that

$$Y_k = \left\{ X_i > 0 | \sum_{j=1}^{i} J_j = k \right\},$$

where $Y_k \sim G(y|\theta)$ can be interpreted as the $k^{th}$ day in which there was rain.

Note that

$$\sum_{k=1}^{n_r} y_k = \sum_{i=1}^{n} x_i$$

The test for a fixed change-point at time $c^*$ in the series then assumes

$$X_1, \ldots, X_{c^*} \sim F(p_1, \theta_1),$$

$$X_{c^*+1}, \ldots, X_n \sim F(p_2, \theta_2),$$

and

$$H_0 : p = p_1 = p_2,\ \theta = \theta_1 = \theta_2,$$

$$H_A : not\ H_0.$$

If

$$n_{r1} = \sum_{i=1}^{c^*} J_i, \quad n_{r2} = \sum_{i=c^*+1}^{n} J_i,$$

So that $n_{r1}$ represents the number of days with rain before the change-point and $n_{r2}$ represents the number of days with rain after the change-point then

$$\hat{p} = \frac{n_r}{n}, \quad \hat{p_1} = \frac{n_{r1}}{c^*}, \quad \hat{p_2} = \frac{n_{r2}}{n - c^*}.$$

Finally,

$$\begin{aligned}
\Lambda_{c^*} &= \frac{\sup_{p,\theta} \prod_{i=1}^{n} f(X_i|p,\hat{\theta})}{\sup_{p_1,\theta_1} \prod_{i=1}^{c^*} f(X_i|p_1,\theta_1) \cdot \sup_{p_2,\theta_2} \prod_{i=c^*+1}^{n} f(X_i|p_2,\theta_2)} \\
&= \frac{\prod_{i=1}^{n} f(X_i|\hat{p},\hat{\theta})}{\prod_{i=1}^{c^*} f(X_i|\hat{p}_1,\hat{\theta}_1) \cdot \prod_{i=c^*+1}^{n} f(X_i|\hat{p}_2,\hat{\theta}_2)} \\
&= \frac{(1-\hat{p})^{n-n_r}(\hat{p})^{n_r} \prod_{k=1}^{n_r} g(Y_k|\hat{\theta})}{(1-\hat{p_1})^{c^*-n_{r1}}(\hat{p_1})^{n_{r1}} \prod_{k=1}^{n_{r1}} g(Y_k|\hat{\theta}_1) \cdot (1-\hat{p_2})^{(n-c^*)-n_{r2}}(\hat{p_2})^{n_{r2}} \prod_{k=n_{r1}+1}^{n} g(Y_k|\hat{\theta}_2)}.
\end{aligned}$$

Importantly,

$$\begin{aligned}
\log(\Lambda_{c^*}) &= (n-n_r)\log(1-\hat{p}) + n_r \log(\hat{p}) + \sum_{k=1}^{n_r} \log g(Y_k|\hat{\theta}) \\
&\quad -(n-n_{r1})\log(1-\hat{p}_1) - n_{r1}\log(\hat{p}_1) - \sum_{k=1}^{n_{r1}} \log g(Y_k|\hat{\theta}_1) \\
&\quad -(n-n_{r2})\log(1-\hat{p}_2) - n_{r2}\log(\hat{p}_2) - \sum_{k=n_{r1}+1}^{n_r} \log g(Y_k|\hat{\theta}_2).
\end{aligned}$$

Let

$$\mathbf{y}_1 = \{Y_k\}_{k=1}^{n_{r1}}, \quad \mathbf{y}_2 = \{Y_k\}_{k=n_{r1}+1}^{n_r},$$

so that $\mathbf{y}_1$ can be interpreted as the series of rain intensities when there was rain before the change-point, and $\mathbf{y}_2$ can be interpreted likewise after the change-point. Also, we write the log-likelihood for $\mathbf{y}$ as

$$L(\mathbf{y}|\hat{\theta}) = \sum_{k=1}^{n_r} \log g(Y_k|\hat{\theta}),$$

where $\mathbf{y}_1$ and $\mathbf{y}_2$ are similarly defined. Therefore,

$$\begin{aligned}
\log(\Lambda_{c^*}) &= (n-n_r)\log\left(1-\frac{n_r}{n}\right)+n_r\log\left(\frac{n_r}{n}\right) \\
&\quad -(n-n_{r1})\log\left(1-\frac{n_{r1}}{c^*}\right)-n_{r1}\log\left(\frac{n_{r1}}{c^*}\right) \\
&\quad -(n-n_{r2})\log\left(1-\frac{n_{r2}}{n-c^*}\right)-n_{r2}\log\left(\frac{n_{r2}}{n-c^*}\right) \\
&\quad -L(\mathbf{y}|\hat{\theta})-L(\mathbf{y}_1|\hat{\theta}_1)-L(\mathbf{y}_2|\hat{\theta}_2).
\end{aligned}$$

## 4.2 $MExp$ and the LRT

If $X_1, X_2, \ldots, X_n \sim MExp(p,\lambda)$, then $Y_1, Y_2, \ldots, Y_{n_r} \sim Exp(\lambda)$

$$\begin{aligned}
L(\mathbf{y}|\hat{\lambda}) &= \sum_{k=1}^{n_r}\log g(Y_k|\hat{\lambda}) = \sum_{k=1}^{n_r}\log\left(\hat{\lambda}e^{-\hat{\lambda}y_k}\right) \\
&= \sum_{k=1}^{n_r}\log\hat{\lambda}-\hat{\lambda}\sum_{k=1}^{n_r}y_k = n_r\log\hat{\lambda}-\hat{\lambda}\sum_{k=1}^{n_r}y_k.
\end{aligned}$$

Since

$$\hat{\lambda}=\frac{n_r}{\sum_{k=1}^{n_r}y_k},$$

then

$$L(\mathbf{y}|\hat{\lambda}) = n_r \log\left(\frac{n_r}{\sum_{k=1}^{n_r} y_k}\right) - \frac{n_r}{\sum_{k=1}^{n_r} y_k}\left(\sum_{k=1}^{n_r} y_k\right)$$

$$= n_r \log n_r - n_r \log \sum_{k=1}^{n_r} y_k - n_r.$$

Because

$$\hat{\lambda}_1 = \frac{n_{r1}}{\sum_{k=1}^{n_{r1}} y_k}, \quad \hat{\lambda}_2 = \frac{n_{r2}}{\sum_{k=n_{r1}+1}^{n_r} y_k},$$

we can write

$$L(\mathbf{y}_1|\hat{\lambda}_1) = n_{r1} \log n_{r1} - n_{r1} \log \sum_{k=1}^{n_{r1}} y_k - n_{r1},$$

$$L(\mathbf{y}_2|\hat{\lambda}_2) = n_{r2} \log n_{r2} - n_{r2} \log \sum_{k=n_{r1}+1}^{n_r} y_k - n_{r2}.$$

Because $n_r = n_{r1} + n_{r2}$,

$$L(\mathbf{y}|\hat{\theta}) - L(\mathbf{y}_1|\hat{\theta}_1) - L(\mathbf{y}_2|\hat{\theta}_2)$$

$$= n_r \log n_r - n_{r1} \log n_{r1} - n_{r2} \log n_{r2}$$

$$
\begin{aligned}
L(\mathbf{y}|\hat{\theta}) - L(\mathbf{y}_1|\hat{\theta}_1) - L(\mathbf{y}_2|\hat{\theta}_2) &= n_r \log n_r - n_{r1} \log n_{r1} - n_{r2} \log n_{r2} \\
&\quad - n_r \log \sum_{k=1}^{n_r} y_k + n_{r1} \log \sum_{k=1}^{n_{r1}} y_k + n_{r2} \log \sum_{k=n_{r1}+1}^{n_r} y_k \\
&= n_r \log n_r - n_{r1} \log n_{r1} - n_{r2} \log n_{r2} \\
&\quad + n_{r1} \log \sum_{k=1}^{n_{r1}} y_k - n_{r1} \log \sum_{k=1}^{n_r} y_k \\
&\quad + n_{r2} \log \sum_{k=n_{r1}+1}^{n_r} y_k - n_{r2} \log \sum_{k=1}^{n_r} y_k.
\end{aligned}
$$

Define

$$
T^* = \frac{\sum_{k=1}^{n_{r1}} y_k}{\sum_{k=1}^{n_r} y_k} \Rightarrow 1 - T^* = \frac{\sum_{k=n_{r1}+1}^{n_r} y_k}{\sum_{k=1}^{n_r} y_k},
$$

then

$$
\begin{aligned}
L(\mathbf{y}|\hat{\theta}) - L(\mathbf{y}_1|\hat{\theta}_1) - L(\mathbf{y}_2|\hat{\theta}_2) &= n_r \log n_r - n_{r1} \log n_{r1} - n_{r2} \log n_{r2} \\
&\quad + n_{r1} \log T^* + n_{r2} \log(1 - T^*).
\end{aligned}
$$

Finally, when testing for a fixed change-point at $c^*$ assuming an $MExp$ distribution,

$$\begin{aligned}
\log(\Lambda_{c^*}) &= (n-n_r)\log\left(1-\frac{n_r}{n}\right)+n_r\log\left(\frac{n_r}{n}\right) \\
&\quad -(n-n_{r1})\log\left(1-\frac{n_{r1}}{c^*}\right)-n_{r1}\log\left(\frac{n_{r1}}{c^*}\right) \\
&\quad -(n-n_{r2})\log\left(1-\frac{n_{r2}}{n-c^*}\right)-n_{r2}\log\left(\frac{n_{r2}}{n-c^*}\right) \\
&\quad +n_r\log n_r - n_{r1}\log n_{r1} - n_{r2}\log n_{r2} + n_{r1}\log T^* + n_{r2}\log(1-T^*).
\end{aligned}$$

## 4.3 Rejection Values

Now that we can calculate the likelihood ratio statistic when testing for a change-point in a random sample of $MExp$ observations, it is necessary to calculate the proper values at which the null hypothesis will be rejected. Commonly, it is said that the likelihood test rejects for small values; however, what exactly are the aforementioned values? If it is desired for the test to have a $(1-\alpha)\cdot 100\%$ confidence level, we seek a value $k$ such that $P(\Lambda_{c^*} < k) = \alpha$. If the $k$ is properly chosen, and if we observe that $\Lambda_{c^*} < k$, then according to theory , the likelihood of the null hypothesis being true is small enough in relation to that of the alternative so that we can reject the null hypothesis with $(1-\alpha)\cdot 100\%$ confidence. Often, the statistic that is used is $-2\log(\Lambda_{c^*})$, which is large when $\Lambda_{c^*}$ is small.

In order to find the proper $k$, the distribution of $\Lambda_{c^*}$ under the null hypothesis must be known. Note that if the distribution of $\Lambda_{c^*}$ is not known, but it is observed that $\Lambda_{c^*} = g(T)$, where the distribution of $T$ is known, then the value of $k$ that will cause the null hypothesis

to be rejected can be found through transformation: $P(\Lambda_{c^*} < k) = P(g(T) < k) = P(T < g^{-1}(k))$, in the case where $g$ is invertible. Hence, statisticans often seek a $T$ with a known distribution under the null hypothesis instead of attempting to compute the distribution of $\Lambda_{c^*}$ itself.

When testing for change-points, the exact value of the parameter set $\theta$, under the null hypothesis is not known. Hence, the likelihood ratio test is most powerful when $\Lambda_{c^*}$ (or $T$) is invariant of the unknown parameters. In the case of testing for a change-point in a series of $MExp$ distributed data, $\Lambda_{c^*}$ depends on $\lambda$ through the values $\{x_i\}$, and $\Lambda_{c^*}$ only depends upon $\{x_i\}$ through $T^* = \frac{\sum_{i=1}^{c^*} x_i}{\sum_{i=1}^{n} x_i}$. It is easy to show that $T^* \sim Beta(n_{r1}, n_{r2})$, and thus $\Lambda_{c^*}$ is invariant of $\lambda$.

However, $\Lambda_{c^*}$ is not invariant of the paramter, $p$. A simulation was run to see just how much 95% cut-off changes as value of $p$ changes under the null hypothesis. For $n = 500, c^* = 250$ and for each value of p listed, 1,000,000 values of -2log($\Lambda_{c^*}$) were simulated, and the corresponding 95% cut-off is shown in Figure 4.1.

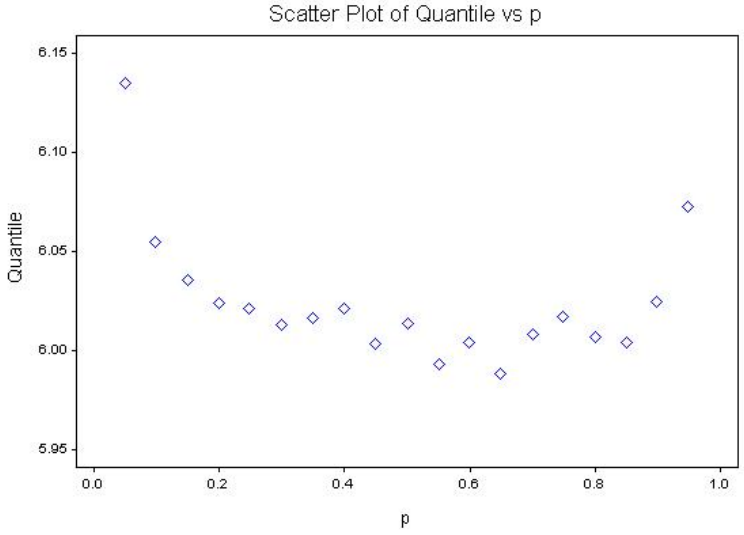


Figure 4.1: $p$ vs. 95% quantile of $\Lambda_{c^*}$ for $n = 500, c^* = 250$

Noticably, the quantiles for rejecting change depend upon the actual value of $p$. However, the quantiles are never significantly different for various $p$, and aside from extreme values of $p$, the 95% cut-off value is consistent. This simulation was run for $n = 500$; so does $\Lambda_{c^*}$ still depend upon $p$ for much larger values of $n$? The graph in Figure 4.2 shows corresponding simulated quantiles when $n = 5000$.

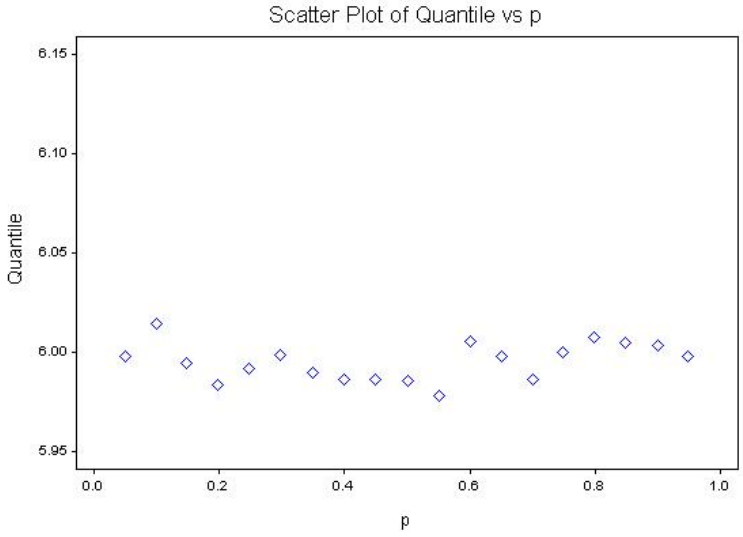


Figure 4.2: $p$ vs. 95% quantile of $\Lambda_{c^*}$ for $n = 5000$, $c^* = 2500$

The dependence of $\Lambda_{c^*}$ on $p$ seems to have completely disappeared, and the 95% quantile seems to be approaching 6.00 asymptotically.

To confirm this observation, the following theorem, Casella & Berger (2002), is introduced.

**Theorem 1** *Let $X_1, \ldots, X_n = \boldsymbol{X}$ be a random sample from a pdf or pmf $f(x|\theta)$. Under regularity conditions and under the null hypothesis ($\theta \in \Theta$), then the distribution of the statistic, $-2\log(\lambda(\boldsymbol{X}))$, where $\lambda(\boldsymbol{X})$ represents the likelihood ratio statistic of the test, converges to a chi squared distribution as $n \to \infty$. The degrees of freedom, d, of the limiting chi squared distribution is given by the difference between the number of free parameters under*

*the alternative hypothesis and the number of free parameters under the null hypothesis.*

When testing for a fixed change-point, $d$ in Theorem 1 always equals the number of free parameters under $H_0$.

We note that $\chi^2_{2,.05} = 5.991465$, which is consistent with the 95% quantiles observed from the simulation.

Of course, the age old statistical quandry finds itself prudent once again: how big does $n$ need to be? The graph below seeks to examine the rate at which $\Lambda_{c^*}$ converges to a chi-squared distribution. If the null hypothesis is rejected when $\Lambda_{c^*} > 5.991465$ (A target Type I error of 95%), about what is the actual Type I error for various finite values of $n$? The graph in Figure 4.3 plots $n$ against the percent of simulated samples in which $\Lambda_{n/2} > 5.991465$; this gives simulated estimation for $\alpha$ that will be referred to as $\hat{\alpha}$.

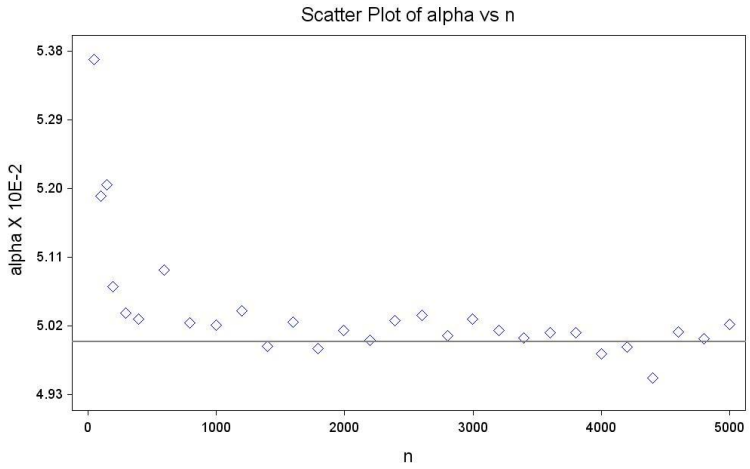


Figure 4.3: $n$ vs. $\hat{\alpha}$ for $c^* = n/2$

Even for sample sizes as low as 50, the actual Type I error is acceptably close to the target Type I error. Once the sample size exceeds 300, the chi-squared quantile appears to apply.

# Chapter 5

# Testing for Unknown Change-Points

The problem of testing for a known change-point is well understood, and the test is relatively easy to execute. However, when it becomes necessary to search for an unknown change-point over a wide range of time, the test becomes significantly more difficult. Recall the previously stated hypothesis for the test of an unknown change-point, $c$, for $l \leq c \leq h$:

$$H_0 : \theta_1 = \ldots = \theta_n$$

$$H_A : \text{there is some integer } c,\ l \leq c \leq h,\ \text{such that}$$

$$\theta_1 = \ldots = \theta_c \neq \theta_{c+1} = \ldots = \theta_n$$

It is intuitive to think that the optimal test statistic when testing for an unknown change-point would be to first calculate $\Lambda_c$ for each possible change-point and then use (as the statistic for unknown change-point test) the one that gives the strongest indication of a

change-point. However, does the mathematical calculation of the likelihood ratio statistic for the test for an unknown change-point confirm the intuition?

When the likelihood ratio is written for this test, the unknown change-point $c$, must be thought of as a parameter. The estimated value of $c$, which is denoted $\hat{c}$, is the change-point which maximizes the likelihood of the alternative hypothesis. The likelihood ratio statistic for this test will be denoted $\Lambda$.

$$
\begin{aligned}
\Lambda &= \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})} \\
&= \frac{\max_{\{\lambda,p\}} L(\lambda, p|\mathbf{x})}{\max_{\{c,\lambda_1,p_1,\lambda_2,p_2\}} L(c, \lambda_1, p_1, \lambda_2, p_2|\mathbf{x})} \\
&= \frac{\max_{\{\lambda,p\}} \prod_{i=1}^{n} f(X_i; p, \lambda)}{\max_c \left(\max_{\{\lambda_1,p_1\}} \prod_{i=1}^{c} f(X_i; p_1, \lambda_1) \cdot \max_{\{\lambda_2,p_2\}} \prod_{i=c+1}^{n} f(X_i; p_2, \lambda_2)\right)} \\
&= \frac{\prod_{i=1}^{n} f(X_i; \hat{p}, \hat{\lambda})}{\max_{l\le c\le h} \left(\prod_{i=1}^{c} f(X_i; \hat{p}_1, \hat{\lambda}_1) \cdot \prod_{i=c+1}^{n} f(X_i; \hat{p}_2, \hat{\lambda}_2)\right)} \\
&= \min_{l\le c\le h} \left\{ \frac{\prod_{i=1}^{n} f(X_i; \hat{p}, \hat{\lambda})}{\prod_{i=1}^{c} f(X_i; \hat{p}_1, \hat{\lambda}_1) \cdot \prod_{i=c+1}^{n} f(X_i; \hat{p}_2, \hat{\lambda}_2)} \right\} \\
&= \min_{l\le c\le h} \{\Lambda_c\}, \\
&\Rightarrow -2\log(\Lambda) = \max_{l\le c\le h} \{-2\log(\Lambda_c)\}.
\end{aligned}
$$

The calculation of $\Lambda$ confirms the intuition.

Also, $\hat{c}$ is defined as the estimated change-point:

$$\hat{c} = \{c|\Lambda_c = \Lambda\}\,.$$

## 5.1 The Distribution of $\hat{c}$ Under $H_0$

The null hypothesis states that there is no change-point, so if $H_0$ is, in fact, true, then it is intuitive to believe that if we were to estimate a change-point, $\hat{c}$, then the distribution of $\hat{c}$ would be uniform over the range of possible change-points.

For $n = 200$, 50,000 values of $\hat{c}$ were calculated, and the plot in Figure 5.1 gives the number of occurences of $\hat{c}$ at each possible value of $c, 10 \leq c \leq 190$.

Obviously, $\hat{c}$ under $H_0$ does not have a uniform distribution, as we would have hoped. Frequently, statisticians notice this anomaly and then proceed to conveniently ignore it. Therefore, is this lack of uniformity that we have just observed really all that important? How does it affect the test? Can it be corrected for? After observing this non-uniformity, now the intuituion is that the probability of a Type I error ($\alpha$) is going to increase if a change-point is estimated as occuring near one of the end-points of the interval. Once again, 50,000 values of $\Lambda$ were simulated under $H_0$ (no change-point) and 95% of these values were less than 12.35583. The graph in Figure 5.2 plots $\hat{c}$ vs the probablility that $\Lambda$ is less than 12.35583 given $\hat{c}$.

Clearly, the conditional probability of a Type I error decreases as $\hat{c}$ gets closer to the end-points. This decrease is never drastic and is only noticeable for $\hat{c}$ that occur right by one of the end-points. Our intuition is wrong again. However, our intuition was right when

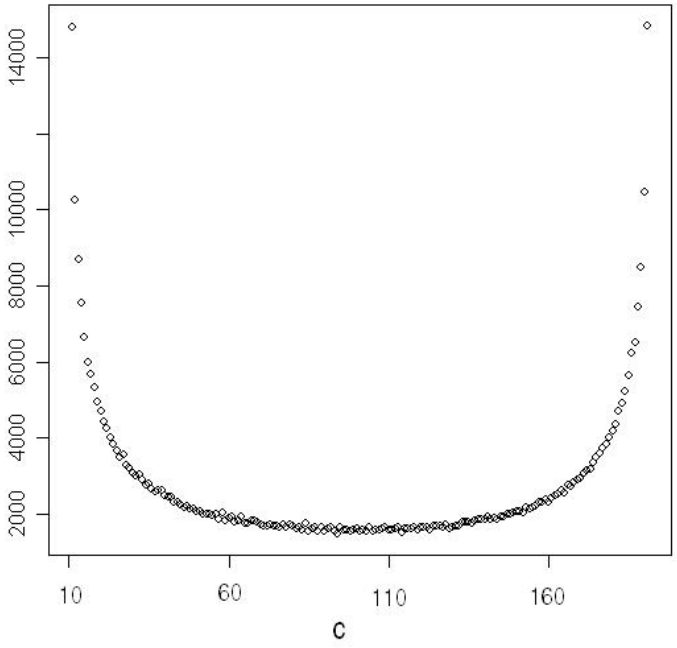


Figure 5.1: A Simulated Mass Function of $\hat{c}$ Under $H_0$

we needed it to be (it confirmed that $\Lambda$ is the likelihood ratio statistic when testing for an unknown change-point) and it was wrong when it could have shown that the U-shape in the distribution of $\hat{c}$ has an adverse effect on the test.

If we determined that this U-shape simply could not be tolerated, we could attempt to correct for it by finding some function of $c$, $q(c)$ such that if

$$\Lambda = \min_{l \leq c \leq h} \left\{ \frac{\Lambda_c}{q(c)} \right\}$$

and

$$\hat{c} = \{c | \Lambda_c = \Lambda\}$$

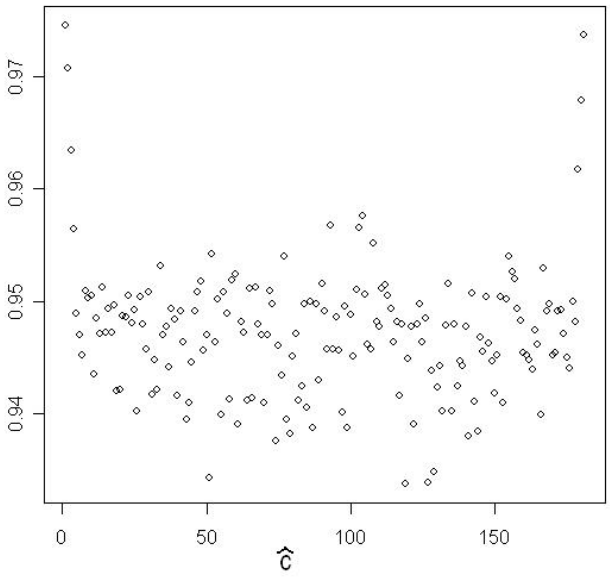


Figure 5.2: $\hat{c}$ vs. $P(\Lambda < 12.35583|\hat{c})$

then finally $\hat{c}$ has a uniform distribution.

## 5.2 Asymptotic Distribution of $\Lambda$

Finding the exact distribution of $\Lambda_c$ is, at best, very difficult. Hence, if one aspires to find the exact distribution of $\Lambda$, where $\Lambda$ is the infimum of the sequence of correlated random variables, $\{\Lambda_c\}_{c=1}^{n}$, it won't take long before one reconsiders his or her objectives. Fortunately, $\Lambda_c$ has an asympotic distribution that only depends upon $d$, the number of free parameters under the null hypothesis. And even more fortunately, $\Lambda$ also has an asymptotic distribution. As $n \to \infty$, the process, $\{\Lambda_c\}_{c=1}^{n}$ behaves as a simple transform of a Brownian bridge. In order to state this result as a theorem we introduce some notation.

Let

$$V_n(t) = \begin{cases} 0 & \text{if } 0 \leq t < 1/(n+1), \\ -2\log\Lambda_{[(n+1)t]} & \text{if } 1/(n+1) \leq t < n/(n+1), \\ 0 & \text{if } n/(n+1) \leq t \leq 1, \end{cases}$$

and

$$B^{(d)}(t) = \sum_{1 \leq i \leq d} B_i^2(t)$$

where $\{B_1(t), 0 \leq t \leq 1\}, \ldots, \{B_d(t), 0 \leq t \leq 1\}$ are independent Brownian bridges. The process, $\{B(t), 0 \leq t \leq 1\}$ is called a Brownian bridge if $B$ is a continuous Gaussian process with $E[B(t)] = 0$ and $E[B(t)B(s)] = \min(t,s) - ts$.

**Theorem 2 (Csorgo & Horvath (1997))** *Under the null hypothesis,*

$$\sup_{0 \leq t \leq 1} t(1-t)V_n(t) \xrightarrow{D} \sup_{0 \leq t \leq 1} B^{(d)}(t)$$

This theorem is an implication of the observation that as $n \longrightarrow \infty$, the process $\{\Lambda_c\}_{c=1}^n$ behaves like a sum of independent Brownian bridges scaled by $t(1-t)$, $(B^{(d)}(t)/t(1-t))$. The above theorem holds under a null hypothesis that assumes any distribution that meets certain regularity conditions, and where $d$ represents the number of free parameters in the null hypothesis.

Unfortunately, as $n$ gets larger, $V_n(t)$ is effectively "blowing up" at the end-points, 0 and 1. This means that

$$\max_{1 \leq c \leq n} \{-2\log(\Lambda_c)\} \to \infty$$

as $n \to \infty$, which is because

$$\lim_{t\to 0} \frac{B^{(d)}(t)}{t(1-t)} = \infty, \text{ and } \lim_{t\to 1} \frac{B^{(d)}(t)}{t(1-t)} = \infty$$

with probability 1.

The U-Shaped distribution of $\hat{c}$ (which showed that under $H_0$ the estimated change-point $\hat{c}$ is more likely to occur near the end-points) is a result of the above phenomenon.

Fortunately,

$$\sup_{a\le t\le 1-b} \left\{ \frac{B^{(d)}(t)}{t(1-t)} \right\} < \infty$$

with probability 1 as long as $0 < a < 1 - b < 1$. Therefore it is possible to determine that

$$P\left\{ \sup_{a\le t\le 1-b} \left( \frac{B^{(d)}(t)}{t(1-t)} \right)^{1/2} \ge x \right\} = \frac{x^d e^{-x^2/2}}{2^{d/2}\Gamma(d/2)} \left\{ \left(1 - \frac{d}{x^2}\right) \log\left( \frac{(1-a)(1-b)}{ab} \right) + \frac{4}{x^2} + O\left(\frac{1}{x^4}\right) \right\}$$

Finally, for the hypothesis

$$H_0 : \theta_1 = \ldots = \theta_n$$

$$H_A : \text{there is some integer c, } l \le c \le h, \text{ such that}$$

$$\theta_1 = \ldots = \theta_c \neq \theta_{c+1} = \ldots = \theta_n$$

Where there are $d$ free parameters under $H_0$ ($\theta_i$ is a vector of dimension $d$ for all $i$). The test statistic, $T$, is

$$T = -2\log(\Lambda) = \max_{l \le c \le h} \{-2\log(\Lambda_c)\}$$

And the approximation

$$P\{T \ge x\} \approx \frac{x^{d/2}e^{-x/2}}{2^{d/2}\Gamma(d/2)} \left\{ \left(1 - \frac{d}{x}\right) \log\left( \frac{\left(1-\frac{l}{n}\right)\left(\frac{h}{n}\right)}{\left(\frac{l}{n}\right)\left(1-\frac{h}{n}\right)} \right) + \frac{4}{x} \right\} := \xi^{(d)}_{[l/n,h/n]}(x)$$

for $x > 0$, can be used to find an approximated P-Value for the test.

If one wishes to find an $(1-\alpha)\cdot 100\%$ level cut-off for the test, one should solve for the value $x^*$ such that under $H_0$

$$\alpha = P\left\{ \max_{l \le c \le h} [-2\log(\Lambda_c)] \ge x^* \right\} \approx \xi^{(d)}_{[l/n,h/n]}(x^*) := \breve{\alpha}.$$

## 5.3 Simulated Quantiles and Convergence Rates

If we wish to test for an unknown change-point in a series of precipitation data $\{X_i\}_{i=1}^n$, and we assume that the precipitations follow a $MExp$ distribution, then the hypotheses for the test are:

$$H_0 : X_1, \ldots, X_n \sim MExp(p, \lambda)$$

$$H_A : \text{there is some integer } c,\ l \le c \le h,\ \text{such that}$$

$$X_1, \ldots, X_c \sim MExp(p_1, \lambda_1)$$

$$X_{c+1}, \ldots, X_n \sim MExp(p_2, \lambda_2).$$

For this test, $d = 2$, so if our target Type I error probability is $\breve{\alpha}$, then we find the cut-off value, $x^*$ by solving $\breve{\alpha} = \xi^{(2)}_{[l/n,h/n]}(x^*)$ for $x^*$. We reject $H_0$ if $-2\log(\Lambda) > x^*$. However, since $x^*$ is derived from an asymptotic distribution, the actual probability of Type I error $\alpha$, is not exactly equal to the target probability $\breve{\alpha}$. When applying the test for finite $n$, $\alpha$ is unknown, but can be estimated through simulation (this estimation is denoted $\hat{\alpha}$) in order to approximate how close $\alpha$ is to value its target value.

For example, if we set $\breve{\alpha} = .05$, $l/n = .1$, and $h/n = .9$, and then we solve for $x^*$ such that $\xi^{(2)}_{[.1,.9]}(x^*) = .05$ and find $x^* = 12.42093$. For each value of $n$ listed in the Tables 1-4, 100,000 values of $-2\log(\Lambda)$ were simulated and the values in the column labeled $\hat{x}^*_n$ represent the corresponding 95% quantile for the simulated vector. In Table 5.1, the values in the column labeled $\hat{\alpha}$ represent the percentage of values that are greater than $x^* = 12.42093$. The simulations were generated using $p = .5$ and $\lambda = 1$.

| $n$ | $\hat{x}^*_n$ | $\hat{\alpha}$ |
|---|---|---|
| 250 | 11.86273 | .03994 |
| 500 | 11.96688 | .04188 |
| 750 | 12.02852 | .04210 |
| 1000 | 12.10251 | .04340 |
| 1500 | 12.14854 | .04499 |
| 2000 | 12.13055 | .04453 |
| 2500 | 12.16562 | .04506 |

Table 5.1: $l/n = .1$, $h/n = .9 \Rightarrow x^* = 12.42093$

In order to see how selecting $l$ and $h$ (which govern the amount of data that is cut-off on each end-point) effects the convergence rates, this process is repeated with $l/n = 1 - h/n = .04$ (Table 2) and $l/n = 1 - h/n = .24$ (Table 3). The results in Table 5.2 were generated using 50,000 iterations for each $n$, and the results in Table 5.3 come from 25,000 iterations.

Apparently, the larger the percentage of the data that is cut-off on each end-point when testing for a change-point, the quicker the test statistic, $\Lambda$, converges to its asympotic

| $n$ | $\hat{x}_n^*$ | $\hat{\alpha}$ |
|---|---|---|
| 250 | 12.49720 | .03535 |
| 500 | 12.80629 | .04148 |
| 750 | 12.77893 | .04182 |
| 1000 | 12.85691 | .04232 |
| 1500 | 12.90765 | .04320 |
| 2000 | 12.96672 | .04428 |
| 2500 | 12.97423 | .04502 |

Table 5.2: $l/n = .04$, $h/n = .96 \Rightarrow x^* = 13.2540$

| $n$ | $\hat{x}_n^*$ | $\hat{\alpha}$ |
|---|---|---|
| 250 | 10.58583 | .04188 |
| 500 | 10.73361 | .04464 |
| 750 | 10.61403 | .04256 |
| 1000 | 10.85994 | .04592 |
| 1500 | 10.88514 | .04704 |
| 2000 | 10.89587 | .04768 |
| 2500 | 10.98804 | .04940 |

Table 5.3: $l/n = .24$, $h/n = .76 \Rightarrow x^* = 11.02771$

distribution.

What happens if instead of ignorning a constant percentage of the data on each end as $n$ increases, we ignore a constant amount of data on each end?

In the Table 5.4 below, for each value of $n$, $\hat{x}_n^*$ represents the 95% quantile for a vector of 10,000 simulated values of $\Lambda$. We set $l = 50$ and $h = n - 50$ for each $n$. Hence, the cut-off value $x_n^*$ that satisifies $\xi^{(2)}_{[50,n-50]}(x_n^*) = .05$ is different for each $n$. Again, $\hat{\alpha}$ represents the percentage of values that are greater than $x_n^*$.

In any change-point test, one could use $l = 1$ and $h = n - 1$ and solve $\xi^{(2)}_{[1,n-1]}(x^*) = \alpha$ for $x^*$, which could then be used as a cut-off value that can test for a change-point at any place in the entire series of data. However, this is a bad idea for several reasons. First, the closer $l/n$ is to 0 or $h/n$ is to 1, the more inflated the cut-off $x^*$ becomes. In other words if $l/n$ is very close to 0 or if $h/n$ is very close to 1, then the $x^*$ that solves $\xi^{(2)}_{[l/n,h/n]}(x^*) = \alpha$

| $n$ | $x_n^*$ | $\hat{x}_n^*$ | $\hat{\alpha}$ |
|---|---|---|---|
| 250 | 11.96915 | 10.95867 | .04164 |
| 500 | 12.42093 | 11.98232 | .04156 |
| 750 | 12.63553 | 12.42703 | .04200 |
| 1000 | 12.7735 | 12.66486 | .04152 |
| 1500 | 12.95214 | 13.04258 | .04376 |
| 2000 | 13.06975 | 13.20545 | .04284 |
| 2500 | 13.15668 | 13.36739 | .04348 |

Table 5.4: $l = 50$, $h = n - 50$

will actually be quite a bit larger than the true $(1 - \alpha)100\%$ cut-off.

In the case of change-point testing for precipitation data, it is especially unadvisable to only ignore small amounts of data on each end because if $n_{r1} = 0$ or if $n_{r2} = 0$ then $\Lambda_c$ can not be calculated. In order to test for a change-point at a certain time using the method described in this paper, there must be observed rain on both sides of the suggested change-point. If one is forced to test for a change-point in a series of data in which no rain is observed on one side of the change-point, then one could test for a change-point in the series $\{J_i\}_{i=1}^n$, which represents the occurence of rain on any given day.

An important observation is that for all simulations, $\hat{\alpha} < \breve{\alpha}$. This implies that the change-point test is conservative, and will never have a larger Type I error probability than desired.

# Chapter 6

# Power Analysis

Now that we have analyzed the likelihood of rejecting the null hypothesis when the null hypothesis is true, we shall now examine the chances of accepting the null hypothesis when it should be rejected. In other words, how succesful is the test at detecting change-points when there actually is a change-point? The ability of a test to reject a false null hypothesis is known as its power, where the Type II error, with a probability denoted by $\beta$, refers to the chances of accepting a null hypothesis that should be rejected. Hence, the power of a test equals the quantity $1 - \beta$.

The Type I error probability, $\alpha$, can be computed, or at least reasonably well approximated, for most any statistical test. However, the power of a test typically depends upon just how false the null hypothesis truly is. The power of any test depends upon the values of the unknown parameters under the alternative hypothesis. In the case of precipitation change-point testing, if there is a change of any magnitude in one or both of $p$ or $\lambda$, then the null hypothesis should be rejected. Keep in mind that the change-point could be anywhere! The values of $p$ and $\lambda$ before and after the change-point, as well as the location of the change-point, will all affect the power of the test. Additionally, the power depends upon

the sample size $n$, and the selection of $l$, $h$, and $\alpha$.

In this section we assume that there is a change-point, before which $p = .4$ and $\lambda = 10$. The value of these parameters after the change-point is varied, as is everything else that can be altered and affect the power, which is then approximated using simulation. Also, $p_1$ and $\lambda_1$ refer to the value of $p$ and $\lambda$ respectively before the change point, and $p_2$ and $\lambda_2$ refer to the parameter values after the change point. Unless otherwise stated, $\alpha = .05$. For large $n$, simulations are painfully slow, and accuracy is not quite as important when simulating power; so for each test, power approximations were generated from 1000 iterations.

## 6.1 The Effect of the Magnitude of the Changes in $p$ and $\lambda$

First, let's assume that $p$ stays constant throughout the entire precipitation series ($p = .4$), which has length $n = 2000$. Also assume that there is a change-point half-way through the data at $c = 1000$. We take, $l/n = 1 - h/n = 0.1$. With $\lambda_1 = 10$, Table 6.1 shows how the value of $\lambda_2$ affects the power.

| $\lambda_2$ | $1 - \beta$ |
|---|---|
| 2 | 1.00 |
| 6 | 1.00 |
| 8 | 0.641 |
| 9 | 0.119 |
| 11 | 0.139 |
| 12 | 0.407 |
| 14 | 0.962 |
| 18 | 1.00 |

Table 6.1: $n = 2000$, $c = 1000$, $p_1 = p_2 = .4$, $\lambda_1 = 10$

Now assume that $\lambda$ stays constant throughout the entire series ($\lambda = 10$), which again has length $n = 2000$ with a change-point at $c = 1000$ and with $l/n = 1 - h/n = 0.1$. For $p_1 = .4$, Table 6.2 shows how the value of $p_2$ affects the power.

| $p_2$ | $1-\beta$ |
|---|---|
| .24 | 1.00 |
| .32 | .814 |
| .36 | .196 |
| .38 | .079 |
| .42 | .074 |
| .44 | .207 |
| .48 | .761 |
| .56 | 1.00 |

Table 6.2: $n = 2000$, $c = 1000$, $\lambda_1 = \lambda_2 = 10$, $p_1 = .4$

Clearly the test is not very effective at detecting small changes in the single parameter, however it does consistently locate larger changes.

How well does the test detect the change-point when both parameters change? We let $n = 2000$, $c = 1000$, $p_1 = .4$, $\lambda_1 = 10$ and $l/n = 1 - h/n = 0.1$. The parameters $p_2$ and $\lambda_2$ are allowed to increase and decrease together, and the corresponding power is listed in Table 6.3.

| $p_2$ | $\lambda_2$ | $1-\beta$ |
|---|---|---|
| .28 | 7 | 0.999 |
| .32 | 8 | 0.879 |
| .36 | 9 | 0.229 |
| .38 | 9.5 | 0.105 |
| .42 | 10.5 | 0.365 |
| .44 | 11 | 0.723 |
| .48 | 12 | 0.992 |
| .52 | 13 | 1.000 |

Table 6.3: $n = 2000$, $c = 1000$, $p_1 = .4$, $\lambda_1 = 10$

Now, is the test any more or less powerful if $p_2$ and $\lambda_2$ change in opposite directions (one increases while the other decreases)? Results are shown in Table 6.4.

Once again, the test shows difficulty detecting small changes; however large changes are easily seen. The test also is more likely to find change-points when both parameters change. For instance, when just $p$ increased from .4 to .44, the power of the test is .207,

| $p_2$ | $\lambda_2$ | $1-\beta$ |
|---|---|---|
| .28 | 13 | 1.0000 |
| .32 | 12 | 0.9850 |
| .36 | 11 | 0.66800 |
| .38 | 10.5 | 0.3460 |
| .42 | 9.5 | 0.0640 |
| .44 | 9 | 0.1790 |
| .48 | 8 | 0..862 |
| .52 | 7 | 1.00 |

Table 6.4: $n = 2000$, $c = 1000$, $p_1 = .4$, $\lambda_1 = 10$

and when just $\lambda$ increased from 10 to 11, the power is .139, however when both changes happen simultaneously, the power becomes .723.

## 6.2 The Effect of Sample Size

For $n = 2000$, $c = 1000$, $p_1 = .4$, $\lambda_1 = 10$ and $l/n = 1 - h/n = 0.1$, and for $p_1 = .44$, $\lambda_1 = 9$, which is a relatively small change, the test does not pick up the change the majority of the time. If $c = n/2$ is held fixed, and $n$ is decreased, the test becomes even less effective. However, by the time $n$ is increased to 6000, the test is very capable at detecting change-points, as shown in Table 6.5.

| $n$ | $1-\beta$ |
|---|---|
| 500 | .0946 |
| 1000 | .1701 |
| 1500 | .2492 |
| 2000 | .3466 |
| 3000 | .5371 |
| 4000 | .6950 |
| 5000 | .7970 |
| 6000 | .8802 |

Table 6.5: $c = n/2$, $\lambda_1 = 10$, $\lambda_2 = 9$, $p_1 = .4$, $p_2 = .44$

## 6.3 The Effect of $c$

All the power simulations presented thus far were run for $c = n/2$. Is the test just as capable of finding change-points that are close to the end-points as it is at finding ones that occur in the middle of the data? The results in Table 6.6 are for the baseline case where, $n = 2000$, $\lambda_1 = 10$, $\lambda_2 = 11.5$, $p_1 = .4$, $p_2 = .46$, and $l/n = 1 - h/n = .05$. Since the test is symmetric with respect to $c$ around $c = n/2$, the test need not be performed for $c > n/2$.

| $c$ | $1 - \beta$ |
|---|---|
| 1000 | .6826 |
| 800 | .6572 |
| 600 | .5714 |
| 400 | .4448 |
| 200 | .2214 |
| 100 | .1078 |

Table 6.6: $n = 2000$, $\lambda_1 = 10$, $\lambda_2 = 11.5$, $p_1 = .4$, $p_2 = .46$, $l/n = 1 - h/n = .05$

Clearly, the test becomes significantly less powerful when the change-point is closer to one of the end-points. How drastic must the change be when the change-point is near the edges in order for the test to be consistent (Table 6.7)?

| $p_2$ | $\lambda_2$ | $1 - \beta$ for $c = 100$ | $1 - \beta$ for $c = 200$ | $1 - \beta$ for $c = 1000$ |
|---|---|---|---|---|
| .44 | 9 | .0668 | .1132 | .3282 |
| .48 | 8 | .1770 | .4646 | .9668 |
| .52 | 7 | .5378 | .9264 | 1.000 |
| .56 | 6 | .9040 | .9988 | 1.000 |
| .60 | 5 | .9976 | 1.000 | 1.000 |

Table 6.7: $n = 2000$, $l = n - h = 100$, $p_1 = .4$, $\lambda_1 = 10$

In order for the test to locate change-points that occur very close to one of the end-points, the change must be a drastic one.

## 6.4 The Effect of $l$ and $h$

If the change-point occurs in the middle of the data, the more data that we cut off on the ends, the more powerful the test becomes, as is seen in Table 6.8.

| $l/n = 1 - h/n$ | $1 - \beta$ |
|---|---|
| .01 | .5260 |
| .05 | .5822 |
| .10 | .6184 |
| .15 | .6570 |
| .25 | 1.000 |

Table 6.8: $n = 2000$, $c = 1000$ $\lambda_1 = 10$, $\lambda_2 = 11.6$, $p_1 = .4$, $p_2 = .45$

However, when the change-point is not in the middle of the data, it is advantageous to pick a smaller value of $l/n$ and $1 - h/n$.

## 6.5 The Effect of $\alpha$

Picking a larger value of $\alpha$ will result in a smaller cut-off value ($x^*$), which will cause the power of the test to increase, as is seen in Table 6.9.

| $\alpha$ | $1 - \beta$ |
|---|---|
| .01 | .632 |
| .05 | .847 |
| .10 | .901 |
| .15 | .936 |
| .20 | .944 |
| .25 | .958 |

Table 6.9: $n = 2000$, $c = 1000$ $\lambda_1 = 10$, $\lambda_2 = 9$, $p_1 = .4$, $p_2 = .48$

## 6.6 The Distribution of $\hat{c}$ Under $H_A$

In cases where a test detects a change-point, how accurate is the estimated value of the location of the change-point, $\hat{c}$?

Using 10,000 iterations, for $n = 2000$, $c = 1000$, $p_1 = .4$, $p_2 = .45$, $\lambda_1 = 10$, $\lambda_2 = 12$, and $l/n = .1$, the power was estimated as .7266 Also, the location of each change-point that was found to be significant was recorded, and Figure 6.1 gives a histogram of the locations. Table 6.10 represents the proportion of significant change-points that fell within the given interval.

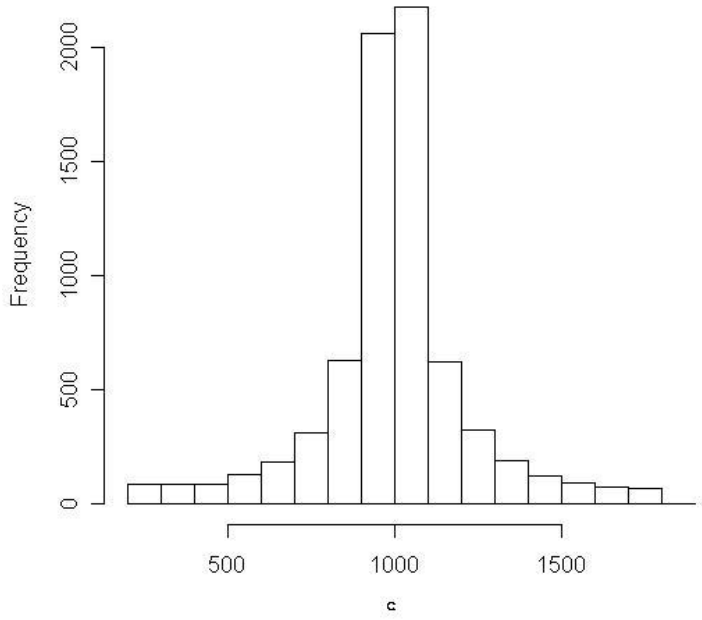


Figure 6.1: The Distribution of $\hat{c}$ Under $H_A$ for $1 - \beta = .7266$

In order to see how the distribution of $\hat{c}$ changes in a case where the power is greater, the process was repeated using 10,000 iterations, for $n = 2000$, $c = 1000$, $p_1 = .4$, $p_2 = .48$, $\lambda_1 = 10$, $\lambda_2 = 12.5$, and $l/n = 1 - h/n = .1$, the power was estimated to be $1 - \beta = .9725$.

| interval | % |
|---|---|
| (900,1100) | 47.58% |
| (800,1200) | 68.04% |
| (700,1300) | 78.52% |
| (600,1400) | 84.93% |

Table 6.10: Interval vs. percent of $\hat{c}$ that fall in the interval, where $1 - \beta = .7266$

As before, Figure 6.2 gives a histogram of the locations, Table 6.11 shows the proportion of significant change-points that fell within the given interval.

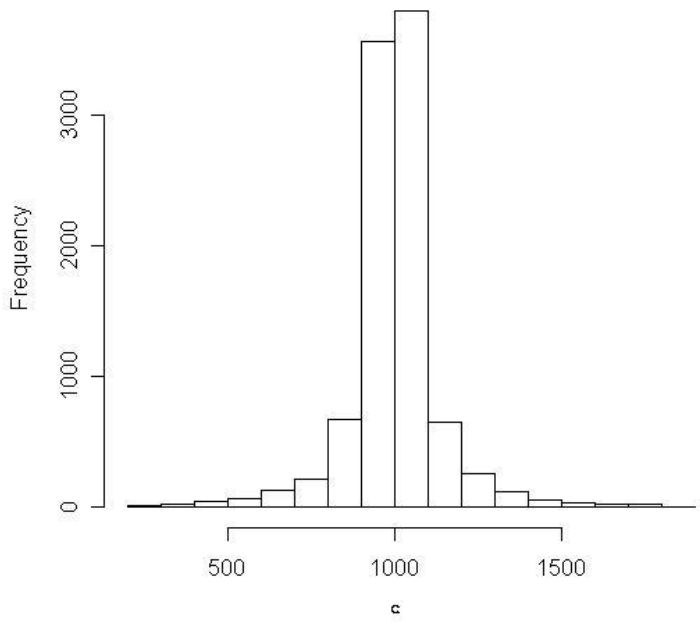


Figure 6.2: The Distribution of $\hat{c}$ Under $H_A$ for $1 - \beta = .9725$

| interval | % |
|---|---|
| (950,1050) | 58.68% |
| (900,1100) | 75.80% |
| (800,1200) | 89.40% |
| (700,1300) | 94.31% |
| (600,1400) | 96.85% |

Table 6.11: Interval vs. percent of $\hat{c}$ that fall in the interval, where $1 - \beta = .9725$

In order to see how the distribution of $\hat{c}$ changes in a case where the power is worse, the process was again repeated using 10,000 iterations, for $n = 2000$, $c = 1000$, $p_1 = .4$, $p_2 = .44$, $\lambda_1 = 10$, $\lambda_2 = 11$, and $l/n = 1 - h/n = .1$, the power was estimated to be $1 - \beta = .3443$. Again, Figure 6.3 gives a histogram of the locations, and then Table 6.12 reports the proportion of significant change-points that fell within the given interval.

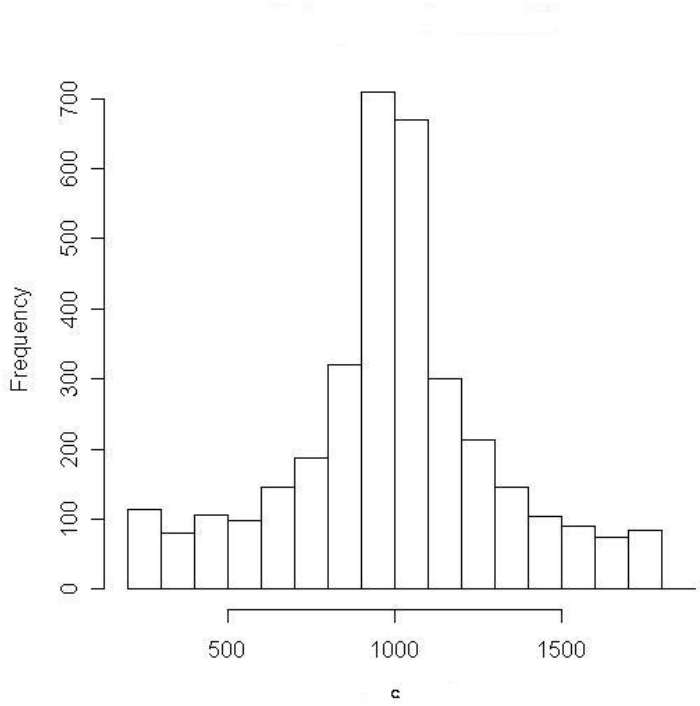


Figure 6.3: The Distribution of $\hat{c}$ Under $H_A$ for $1 - \beta = .3443$

| interval | % |
|---|---|
| (900,1100) | 40.28% |
| (800,1200) | 58.17% |
| (700,1300) | 69.82% |
| (600,1400) | 78.30% |
| (500,1500) | 84.14% |

Table 6.12: Interval vs. percent of $\hat{c}$ that fall in the interval, where $1 - \beta = .3443$

When the test is more powerful, which is usually caused by a more drastic change, the

location of estimated change-points is more likely to be accurate.

# Chapter 7

# Testing Under Gamma Assumptions

The bulk of this change-point study was done assuming that precipitation follows an $MExp$ distribution because it is a reasonable model, the maximum likelihood extimators are easy to find, and simulations run relatively fast when using it. However, most climatologists prefer that rain be modeled using a gamma distribution, whether it be for annual, monthly, or daily data. Therefore, we seek a method of change-point testing for a Type A precipitation model in which $G(z) \sim Gamma(\alpha, \beta)$. In order to apply a likelihood ratio for data that is gamma distributed, we must first develop a method for finding maximum likelihood estimtators for gamma data.

The shape parameter in a gamma distribution is ubiquitously denoted by $\alpha$ throughout statistics and probability literature. Howver, in this paper, $\alpha$ is also used to refer to the probability of Type I error. Henceforth, we will use $\gamma$ to denote the shape parameter in a gamma distribution and $\alpha$ to represent a test's Type I error probability.

## 7.1 Gamma MLEs

If $X_1, X_2, \ldots, X_n \sim Gamma(\gamma, \beta)$ are *iid*, then each $X_i$ has the density function,

$$f(x|\gamma,\beta) = \frac{1}{\Gamma(\gamma)\beta^\gamma} x^{\gamma-1} e^{-x/\beta}, \quad 0 \le x < \infty, \quad \gamma, \beta > 0.$$

The likelihood function becomes

$$\begin{aligned} f_n(\mathbf{x}|\gamma,\beta) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\gamma)\beta^\gamma} x_i^{\gamma-1} e^{-x_i/\beta} \\ &= \frac{1}{(\Gamma(\gamma))^n \beta^{n\gamma}} \prod_{i=1}^{n} \left( x_i^{\gamma-1} e^{-x_i/\beta} \right) \\ &= (\Gamma(\gamma))^{-n} \beta^{-n\gamma} \prod_{i=1}^{n} \left( x_i^{\gamma-1} \right) e^{-\sum_{i=1}^{n} x_i/\beta}. \end{aligned}$$

The log-likelihood is therefore

$$\begin{aligned} L(\mathbf{x}|\gamma,\beta) &= \log(f_n(\mathbf{x}|\gamma,\beta)) = \sum_{i=1}^{n} \log f(x_i|\gamma,\beta) \\ &= -n\gamma \log(\beta) - n\log(\Gamma(\gamma)) + (\gamma-1)\sum_{i=1}^{n} \log(x_i) - \frac{1}{\beta}\sum_{i=1}^{n} x_i. \end{aligned}$$

The MLEs for $\gamma$ and $\beta$ are values of $\gamma$ and $\beta$ for which $\frac{\partial}{\partial\gamma} L(\mathbf{x}|\gamma,\beta) = 0$ and $\frac{\partial}{\partial\beta} L(\mathbf{x}|\gamma,\beta) = 0$. Differentiating yields

$$\frac{\partial}{\partial\beta} L(\mathbf{x}|\gamma,\beta) = -\frac{n\gamma}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i = \frac{1}{\beta}\left( -n\gamma + \frac{1}{\beta}\sum_{i=1}^{n} x_i \right)$$

$$\frac{\partial}{\partial\hat{\beta}} L(\mathbf{x}|\hat{\gamma},\hat{\beta}) = 0 \Rightarrow -n\hat{\gamma} + \frac{1}{\beta}\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \hat{\beta} = \frac{1}{n\hat{\gamma}} \sum_{i=1}^{n} x_i$$

and

$$\frac{\partial}{\partial \gamma} L(\mathbf{x}|\gamma, \beta) = -n \log(\beta) - n \frac{\partial}{\partial \gamma} \log(\Gamma(\gamma)) + \sum_{i=1}^{n} \log(x).$$

The function $\frac{\partial}{\partial \gamma} \log(\Gamma(\gamma))$ is known as the digamma function, and is denoted by $\psi(\gamma)$. Therefore

$$\frac{\partial}{\partial \hat{\gamma}} L(\mathbf{x}|\hat{\gamma}, \hat{\beta}) = 0 \Rightarrow -n \log(\hat{\beta}) - n\psi(\hat{\gamma}) + \sum_{i=1}^{n} \log(x) = 0.$$

$$\Rightarrow -n \log\left(\frac{1}{n\hat{\gamma}} \sum_{i=1}^{n} x_i\right) - n\psi(\hat{\gamma}) + \sum_{i=1}^{n} \log(x) = 0.$$

$$\Rightarrow -n \log\left(\frac{1}{\hat{\gamma}}\right) - n \log\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) - n\psi(\hat{\gamma}) + \sum_{i=1}^{n} \log(x) = 0.$$

$$\Rightarrow \log(\hat{\gamma}) - \psi(\hat{\gamma}) - \log\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) + \frac{1}{n} \sum_{i=1}^{n} \log(x) = 0.$$

This is not an equation that can be solved explicitly. The MLE $\hat{\gamma}$, must be found numerically. However, any software package with a built-in digamma function and non-linear equation solver should be able to solve the above equation for the proper $\hat{\alpha}$.

Also, using any software package with a built-in gamma function and an optimize function (such as R), one could simply maximize

$$L(\mathbf{x}|\gamma) = n\hat{\gamma}\log(\hat{\gamma}) - n\hat{\alpha}\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - n\log(\Gamma(\hat{\gamma})) + (\hat{\gamma}-1)\sum_{i=1}^{n}\log(x_i) - n\hat{\gamma}$$

with respect to $\hat{\gamma}$ to find the maximum likelihood estimate for $\gamma$.

## 7.2 Change-Point Testing for Annual and Monthly Rainfall Data

If $\{X_i\}_{i=1}^n$ is a series of non-zero monthly or annual precipitation data, usually each $X_i$ is thought to be independent with a $Gamma(\gamma, \beta)$ distribution. For a fixed change-point $c^*$, it is assumed that

$$X_1, \ldots, X_{c^*} \sim Gamma(\gamma_1, \beta_1)$$

$$X_{c^*+1}, \ldots, X_n \sim Gamma(\gamma_2, \beta_2)$$

and

$$H_0 : \gamma = \gamma_1 = \gamma_2,\ \beta = \beta_1 = \beta_2$$

$$H_A : \text{not } H_0$$

Thus the likelihood ratio statistic for this test is

$$\Lambda_{c^*} = \frac{\sup_\theta \prod_{i=1}^{n} f(X_i|\theta)}{\sup_{\theta_1} \prod_{i=1}^{c^*} f(X_i|\theta_1) \cdot \sup_{\theta_2} \prod_{i=c^*+1}^{n} f(X_i|\theta_2)}$$

$$= \frac{\prod_{i=1}^{n} f(X_i|\hat{\gamma},\hat{\beta})}{\prod_{i=1}^{c^*} f(X_i|\hat{\gamma}_1,\hat{\beta}_1)\cdot\prod_{i=c^*+1}^{n} f(X_i|\hat{\gamma}_2,\hat{\beta}_2)}.$$

Hence

$$\begin{aligned}
\log(\Lambda_{c^*}) &= \sum_{i=1}^{n}\log f(X_i|\hat{\gamma},\hat{\beta}) - \sum_{i=1}^{c^*}\log f(X_i|\hat{\gamma}_1,\hat{\beta}_1) - \sum_{i=c^*+1}^{n}\log f(X_i|\hat{\gamma}_2,\hat{\beta}_2) \\
&= n\hat{\alpha}\log(\hat{\gamma}) - n\hat{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) - n\log(\Gamma(\hat{\gamma})) + (\hat{\gamma}-1)\sum_{i=1}^{n}\log(x_i) - n\hat{\gamma} \\
&\quad -c^*\hat{\gamma}_1\log(\hat{\gamma}_1) + c^*\hat{\gamma}_1\log\left(\frac{1}{c^*}\sum_{i=1}^{c^*}x_i\right) + c^*\log(\Gamma(\hat{\gamma}_1)) - (\hat{\gamma}_1-1)\sum_{i=1}^{c^*}\log(x_i) + c^*\hat{\gamma}_1 \\
&\quad -(n-c^*)\hat{\gamma}_2\log(\hat{\gamma}_2) + (n-c^*)\hat{\gamma}_2\log\left(\frac{1}{(n-c^*)}\sum_{i=c^*+1}^{n-c^*}x_i\right) + (n-c^*)\log(\Gamma(\hat{\gamma}_2)) \\
&\quad -(\hat{\gamma}_2-1)\sum_{i=c^*+1}^{n-c^*}\log(x_i) + (n-c^*)\hat{\gamma}_2,
\end{aligned}$$

where $\hat{\gamma},\hat{\beta}$ are the MLE's for the series $\{X_i\}_{i=1}^{n}$,

$\hat{\gamma}_1,\hat{\beta}_1$ are the MLE's for the series $\{X_i\}_{i=1}^{c^*}$,

$\hat{\gamma}_2,\hat{\beta}_2$ are the MLE's for the series $\{X_i\}_{i=c^*+1}^{n}$.

$\Lambda_{c^*}$ does have an asymptotic $\chi^2$ distribution with 2 degrees of freedom. So, we would reject the null hypothesis with 95% confidence if $\Lambda_{c^*} > 5.991465$. An example of this test can be seen in Wilks (2006).

When testing for an unknown change-point, the test statistic is, of course,

$$T = \max_{l\le c\le h}\{-2\log(\Lambda_c)\} = -2\log(\Lambda)$$

To find the proper cut-off points for this test, knowing that the null hypothesis has two free parameters, we could solve $\alpha = \xi^{(2)}_{[l/n,h/n]}(x^*)$ to find $x^* = 12.42093$ when $\alpha = .05$ and $l = 1-h = .1$. Naturally, however, monthly and annual rainfall data will have less

observations than daily data. Hence, for smaller values of $n$, Table 7.1 shows a simulated 95% cut-off point, and compares it to the simulated Type 1 error probability ($\hat{\alpha}$) had $x^* = 12.42093$ been used as the cut-off.

| $n$ | $\hat{x}_n^*$ | $\hat{\alpha}$ |
|---|---|---|
| 50 | 12.53647 | .05170 |
| 100 | 12.02759 | .04300 |
| 150 | 12.01029 | .04220 |
| 200 | 11.89313 | .04080 |
| 250 | 11.91596 | .04056 |
| 500 | 12.01977 | .04088 |
| 750 | 12.03077 | .04168 |
| 1000 | 12.05874 | .04252 |

Table 7.1: $l/n = .1$, $h/n = .9 \Rightarrow x^* = 12.42093$

Interestingly, Table 7.1 shows a that the simulated value $\hat{x}_n^*$, is larger for smaller values of $n$, quickly shrinks, and then begins to increases slightly as $n$ increases. This phenonmenon was not observed for any of the simulations involving the $MExp$ distribution, and is due to the fact that when using a gamma density, the distribution of the likelihood ratio, $\Lambda$, depends upon the value of the parameter $\gamma$.

In order to illustrate the dependence of $\Lambda$ on $\gamma$, Table 7.2 shows $\hat{x}_n^*$ and $\hat{\alpha}$ for simulations under various values of $\gamma$ when $n = 50$ and $l/n = 1 - h/n = .1 \Rightarrow x^* = 12.42093$. The process is repeated in Table 7.3 for $n = 250$ and again in 7.4 for $n = 250$. Clearly for small $n$ the dependence on $\gamma$ is fairly drastic, however as $n$ gets larger and the asymptotic distrubtion of $\Lambda$ is reached, the dependence on $\gamma$ disappears.

| $\gamma$ | $\hat{x}_n^*$ | $\hat{\alpha}$ |
|---|---|---|
| .25 | 12.59333 | .05336 |
| .5 | 12.50611 | .05168 |
| 1 | 12.38375 | .04940 |
| 2 | 11.92313 | .04104 |
| 4 | 11.25373 | .02992 |

Table 7.2: $\gamma$ vs. $\hat{x}_n^*$ and $\hat{\alpha}$ for $n = 50$

| $\gamma$ | $\hat{x}_n^*$ | $\hat{\alpha}$ |
|---|---|---|
| .25 | 12.24498 | .04600 |
| .5 | 11.97263 | .04100 |
| 1 | 11.95339 | .03980 |
| 2 | 11.98542 | .04040 |
| 4 | 11.70747 | .03880 |

Table 7.3: $\gamma$ vs. $\hat{x}_n^*$ and $\hat{\alpha}$ for $n = 250$

| $\gamma$ | $\hat{x}_n^*$ | $\hat{\alpha}$ |
|---|---|---|
| .25 | 12.06815 | .04320 |
| .5 | 12.04417 | .04350 |
| 1 | 12.04551 | .04200 |
| 2 | 11.92376 | .04210 |
| 4 | 12.04468 | .04280 |

Table 7.4: $\gamma$ vs. $\hat{x}_n^*$ and $\hat{\alpha}$ for $n = 500$

## 7.3 Validation of $MGam$ as a Model for Daily Rain Data

Convention suggests that if $\{X_i\}_{i=1}^n$ is a daily precipitation series, then each $X_i$ is *iid* with cdf

$$F(x) = 1 - p + pG(x),$$

for $0 \leq p \leq 1$ where $G(x) \sim Gamma(\gamma, \beta)$. Such a distribution will be refered to as an $MGam$ for modified gamma. The mixed *pdf* is defined as

$$g(x|p,\gamma,\beta) = \begin{cases} \frac{p}{\Gamma(\gamma)\beta^{\alpha}} x^{\gamma-1} e^{-x/\beta} & \text{if } x > 0, \\ 1-p & \text{if } x = 0, \\ 0 & \text{else.} \end{cases}$$

Previously, we assumed a $MExp$ distribution because of its relative simplicity. How much information about the behavior of precipitation is being lost by using and exponential model instead of a gamma one? Assume that the series $\{Y_i\}_{i=1}^{2823}$ is observed precipitation on the wet days in our data. Now that we have presented the methods for finding MLE's under gamma assumptions, we can perform the following test:

$$H_0 : Y_1, Y_2, \ldots, Y_{2823} \sim exp(\lambda),$$

$$H_A : Y_1, Y_2, \ldots, Y_{2823} \sim gamma(\gamma, \beta).$$

An exponential model is a specific case of a gamma model, so the likelihood under the alternative must be greater than or equal to the likelihood under the null.

The likelihood ratio statistic is found as follows:

$$\lambda(\mathbf{y}) = \frac{\max_{\lambda} L(\mathbf{y}|\lambda)}{\max_{\gamma,\beta} L(\mathbf{y}|\alpha,\beta)} = \frac{L(\mathbf{y}|\hat{\lambda})}{L(\mathbf{y}|\hat{\gamma},\hat{\beta})}.$$

Therefore,

$$
\begin{aligned}
-2\log(\lambda(\mathbf{y})) &= -2\log(L(\mathbf{y}|\hat{\lambda})) + 2\log(L(\mathbf{y}|\hat{\gamma},\hat{\beta})) \\
&= -2\left(2823\log(\hat{\lambda}) - \hat{\lambda}\sum_{i=1}^{2823} y_i\right) \\
&\quad +2\left(2823\hat{\gamma}\log(\hat{\gamma}) - 2823\hat{\gamma}\log\left(\frac{1}{2823}\sum_{i=1}^{2823} y_i\right)\right) \\
&\quad +2\left(-2823\log(\Gamma(\hat{\gamma})) + (\hat{\gamma}-1)\sum_{i=1}^{2823}\log(y_i) - 2823\hat{\gamma}\right) \\
&= 96.9521,
\end{aligned}
$$

where $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} y_i}$ and $\hat{\gamma}$ and $\hat{\beta}$ are the numerically computed MLE's.

The number of degrees of freedom under the alternative hypothesis minus the number of degrees of freedom under the null is 1, so the asymptotic distribution of this likelihood ratio statistic is $\chi^2_1$. The 95% cut-off is 3.841459, which is significantly less than the computed value $\lambda(\mathbf{y})$. Thus we reject an exponential distribution with a very high degree of certainty. We conclude that it is beneficial to model the rain series under gamma assumptions as opposed to the exponential assumptions presented thus far.

## 7.4 Fixed Change-Point Testing Under $MGam$ Assumptions

If $\{X_i\}_{i=1}^n$ is a series of daily precipitation data, usually each $X_i$ is thought to be independent observations from a $MGam(p,\gamma,\beta)$ distribution. For a fixed change-point at $c^*$, it is assumed that

$$X_1,\ldots,X_{c^*} \sim MGam(p_1,\gamma_1,\beta_1),$$

$$X_{c^*+1}, \ldots, X_n \sim MGam(p_2, \gamma_2, \beta_2),$$

and

$$H_0 : p = p_1 = p_2,\ \gamma = \gamma_1 = \gamma_2,\ \beta = \beta_1 = \beta_2,$$

$$H_A^* : not\ H_0.$$

As in the construction for $\Lambda_{c^*}$ under the assumption of an $MExp$ distribution, let $n_r$ be the total number of wet days, $n_{r1}$ be the number of wet days before the fixed change-point at $c^*$ and let $n_{r2}$ be the number of wet days after the change-point.

Also, let the *iid* series

$$\{Y_i\}_{i=1}^{n} \sim Gamma(\gamma, \beta)$$

represent each wet day in the length-$n$ series, and likewise let

$$\{Y_i\}_{i=1}^{n_{r1}} \sim Gamma(\gamma_1, \beta_1),\ \{Y_i\}_{i=n_{r1}+1}^{n_r} \sim Gamma(\gamma_2, \beta_2)$$

represent the wet days before and after the change-point, respectively, where $n_{r2} = n_r - n_{r1}$. If, as for the general Type A precipitation model,

$$\mathbf{y} = \{Y_i\}_{i=1}^{n_r},\ \mathbf{y}_1 = \{Y_i\}_{i=1}^{n_{r1}},\ \text{and}\ \mathbf{y}_2 = \{Y_i\}_{i=n_{r1}+1}^{n_r},$$

then,

$$\hat{\gamma}, \hat{\beta}\ \text{are the MLE's for the series}\ \mathbf{y},$$

$$\hat{\gamma}_1, \hat{\beta}_1\ \text{are the MLE's for the series}\ \mathbf{y}_1,$$

$$\hat{\gamma}_2, \hat{\beta}_2\ \text{are the MLE's for the series}\ \mathbf{y}_2.$$

These estimates are found using the previously described methods, and the maximized log-likelihood for each of those series of observations is given by

$$L(\mathbf{y}|\hat{\gamma},\hat{\beta}) = -n\hat{\gamma}\log(\hat{\beta}) - n\log(\Gamma(\hat{\gamma})) + (\hat{\gamma}-1)\sum_{i=1}^{n}\log(y_i) - \frac{1}{\hat{\beta}}\sum_{i=1}^{n} y_i,$$

$$L(\mathbf{y}_1|\hat{\gamma}_1,\hat{\beta}_1) = -n_{r1}\hat{\gamma}_1\log(\hat{\beta}_1) - n_{r1}\log(\Gamma(\hat{\gamma}_1)) + (\hat{\gamma}_1-1)\sum_{i=1}^{n_{r1}}\log(y_i) - \frac{1}{\hat{\beta}_1}\sum_{i=1}^{n_{r1}} y_i,$$

$$L(\mathbf{y}_2|\hat{\gamma}_2,\hat{\beta}_2) = -n_{r2}\hat{\gamma}_2\log(\hat{\beta}_2) - n_{r2}\log(\Gamma(\hat{\gamma}_2)) + (\hat{\gamma}_2-1)\sum_{i=1}^{n_{r2}}\log(y_i) - \frac{1}{\hat{\beta}_2}\sum_{i=1}^{n_{r2}} y_i.$$

Now we can compute

$$\begin{aligned}
\Lambda_{c^*} &= \frac{\sup_{p,\gamma,\beta}\prod_{i=1}^{n} f(X_i|p,\gamma,\beta)}{\sup_{p_1,\gamma_1,\beta_1}\prod_{i=1}^{c^*} f(X_i|p_1,\gamma_1,\beta_1)\cdot\sup_{p_2,\gamma_2,\beta_2}\prod_{i=c^*+1}^{n} f(X_i|p_2,\gamma_2,\beta_2)} \\
&= \frac{\prod_{i=1}^{n} f(X_i|\hat{p},\hat{\gamma},\hat{\beta})}{\prod_{i=1}^{c^*} f(X_i|\hat{p}_1,\hat{\gamma}_1,\hat{\beta}_1)\cdot\prod_{i=c^*+1}^{n} f(X_i|\hat{p}_2,\hat{\gamma}_2,\hat{\beta}_2)} \\
&= \frac{(1-\hat{p})^{n-n_r}(\hat{p})^{n_r}\prod_{i=1}^{n_r} f(Y_i|\hat{\gamma},\hat{\beta})}{(1-\hat{p_1})^{c^*-n_{r1}}(\hat{p_1})^{n_{r1}}\prod_{i=1}^{n_{r1}} f(Y_i|\hat{\gamma}_1,\hat{\beta}_1)\cdot(1-\hat{p_2})^{(n-c^*)-n_{r2}}(\hat{p_2})^{n_{r2}}\prod_{i=n_{r1}+1}^{n_r} f(Y_i|\hat{\gamma}_2,\hat{\beta}_2)}.
\end{aligned}$$

So,

$$\begin{aligned}
\log(\Lambda_{c^*}) &= (n-n_r)\log(1-\hat{p}) + n_r\log(\hat{p}) - (c^*-n_{r1})\log(1-\hat{p_1}) \\
&\quad -n_{r1}\log(\hat{p_1}) - ((n-c^*)-n_{r2})\log(1-\hat{p_2}) - n_{r2}\log(\hat{p_2}) \\
&\quad +\sum_{i=1}^{n_r}\log f(Y_i|\hat{\gamma},\hat{\beta}) - \sum_{i=1}^{n_{r1}}\log f(Y_i|\hat{\gamma}_1,\hat{\beta}_1) - \sum_{i=n_{r1}+1}^{n_r}\log f(Y_i|\hat{\gamma}_2,\hat{\beta}_2)
\end{aligned}$$

and

$$\begin{aligned}\log(\Lambda_{c^*}) &= (n-n_r)\log\left(1-\frac{n_r}{n}\right)+n_r\log\left(\frac{n_r}{n}\right)-(c^*-n_{r1})\log\left(1-\frac{n_{r1}}{c^*}\right)\\ &\quad -n_{r1}\log\left(\frac{n_{r1}}{c^*}\right)-((n-c^*)-n_{r2})\log\left(1-\frac{n_{r2}}{n-c^*}\right)-n_{r2}\log\left(\frac{n_{r2}}{n-c^*}\right)\\ &\quad +L(\mathbf{y}|\hat{\gamma},\hat{\beta})-L(\mathbf{y}_1|\hat{\gamma}_1,\hat{\beta}_1)-L(\mathbf{y}_2|\hat{\gamma}_2,\hat{\beta}_2).\end{aligned}$$

For fixed $c^*$, as $n\to\infty$, $-2\log(\Lambda_{c^*})\to\chi^2_3$ where $\chi^2_3$ is a chi-squared random variable with three degrees of freedom (because there are 3 free parameters under $H_0$). Therefore, $H_0$ rejects in favor of $H_A^*$ with 95% confidence if $-2\log(\Lambda_{c^*})>7.8147$.

## 7.5 Testing for an Unknown Change-Point Under $MGam$ Assumptions

As before, when testing for an unknown change-point, we test the same $H_0$ (that there is no change-point) against the alternative, $H_A$, that there is a change-point $c$ such that $l\le c\le h$. The test statistic is

$$T=-2\log(\Lambda)=\max_{l\le c\le h}\{-2\log(\Lambda_c)\}.$$

Because there are three free parameters under $H_0$, for large $n$, $H_A$ rejects if $-2\log(\Lambda)>x^*$ when $\xi^{(3)}_{[l/n,h/n]}(x^*)=\alpha$. For instance, if we ignore 10% of the observations on each end, then because $\xi^{(3)}_{[.1,.9]}(14.92045)=.05$, We reject $H_A$ with 95% confidence if $-2\log(\Lambda)>14.92045$.

# Chapter 8

# Change-Point Testing for Markov Chain Models

All of the change-point testing methods developed thus far in this study have been for Type A precipitation models, in which the series, $\{J_i\}_{i=1}^n$ is assumed to be independent. In order to develop change-point tests for Type B models, in which $\{J_i\}_{i=1}^n$ is considered to behave like a first-order, two-state Markov chain, the methods need to be altered only slightly.

Instead of a single parameter, $p$, to model the occurrence of rain, Type B models use the following two parameters to approximate the occurrence of rain:

$$p_0 = P\left[J_k = 1 | J_{k-1} = 0\right],$$

$$p_1 = P\left[J_k = 1 | J_{k-1} = 1\right].$$

## 8.1 Markov Chain Inference

In order to find the maximum likelihood estimators for $p_0$ and $p_1$, we must first develop some notation for inference of Markov chains, Anderson & Goodman (1957). Let

$$n_{01} = \sum_{i=1}^{n-1} I_i^{(01)} \quad \text{where} \quad I_i^{(01)} = \begin{cases} 1 & \text{if } J_i = 0 \text{ and } J_{i+1} = 1, \\ 0 & \text{else,} \end{cases}$$

$$n_{11} = \sum_{i=1}^{n-1} I_i^{(11)} \quad \text{where} \quad I_i^{(11)} = \begin{cases} 1 & \text{if } J_i = 1 \text{ and } J_{i+1} = 1, \\ 0 & \text{else.} \end{cases}$$

We can interpret, in terms of the precipitation model, $n_{01}$ as the number of days in which it rained when there was no rain the day before and $n_{11}$ as the number of days in which it rained when there was rain on the previous day. Also let

$$n_1 = \sum_{i=1}^{n-1} J_i,$$

$$n_0 = (n-1) - n_1.$$

Now we can find

$$\hat{p}_0 = \frac{n_{01}}{n_0},$$

$$\hat{p}_1 = \frac{n_{11}}{n_1}.$$

## 8.2 The Test for Markov Chain Structure in the Data

The data set has missing values, which is irrelevent when $\{J_i\}_{i=1}^{6353}$ is assumed to be an independent series, however we can still test for a Markov chain structure in that series. So, the following hypotheses are to be tested:

$$H_0\text{: } \{J_i\}_{i=1}^{6353} \text{ in an independent series, where each } J_i \text{ has success probability } p,$$

$$H_A\text{: } \{J_i\}_{i=1}^{6353} \text{ behaves as a first-order, two-state Markov chain.}$$

The likelihood ratio for this test is:

$$\Lambda_{MC} = \frac{\hat{p}^{n_1}(1-\hat{p})^{n_0}}{\hat{p_0}^{n_{01}}(1-\hat{p_0})^{n_0-n_{01}}\hat{p_1}^{n_{11}}(1-\hat{p_1})^{n_1-n_{11}}},$$

where

$$\hat{p} = \frac{n_1}{n-1}.$$

Hence

$$\begin{aligned}\log(\Lambda_{MC}) &= n_1 \log \hat{p} + n_0 \log(1-\hat{p}) - n_{01} \log \hat{p}_0 - (n_0 - n_{01}) \log(1-\hat{p}_0) \\ &\quad -n_{11} \log \hat{p}_1 - (n_1 - n_{11}) \log(1-\hat{p}_1),\end{aligned}$$

$$\Rightarrow -2\log(\Lambda_{MC}) = 20.71256, \text{ where } -2\log(\Lambda_{MC}) \sim \chi^2_2.$$

This test has a 95% cut-off of $\chi^2_2 = 5.991465 < 20.71256$. Therefore, we reject the null hypothesis, and recognize that despite the fact that the data has missing values, there is a noticeable Markov chain tendency. We also found $\hat{p}_0 = .4190$ and $\hat{p}_1 = .4761$.

## 8.3 Markov Chain Change-Point Testing

If we assume that $\{J_i\}_{i=1}^n$ follows a two-state, first-order Markov chain, and we wish to test for a change-point in the process at $c^*$, then we first presume the following:

$$\{J_i\}_{i=1}^{c^*} \text{ has } \mathbf{P}_1 = \begin{pmatrix} (1-p_{0.1}) & p_{0.1} \\ (1-p_{1.1}) & p_{1.1} \end{pmatrix},$$

$$\{J_i\}_{i=c^*+1}^{n} \text{ has } \mathbf{P}_2 = \begin{pmatrix} (1-p_{0.2}) & p_{0.2} \\ (1-p_{1.2}) & p_{1.2} \end{pmatrix}.$$

The hypothesis to be tested promptly becomes,

$$H_0 : \mathbf{P} = \mathbf{P}_1 = \mathbf{P}_2,$$

$$H_A : \mathbf{P}_1 \neq \mathbf{P}_2.$$

In order compute the likelihood ratio statistic, we must first find all maximum likelihood estimators, which requires some more notation. Let

$$n_{01.1} = \sum_{i=1}^{c^*-1} I_i^{(01)},\ n_{01.2} = \sum_{i=c^*}^{n-1} I_i^{(01)},$$

where

$$I_i^{(01)} = \begin{cases} 1 & \text{if } \mathrm{J}_i = 0 \text{ and } J_{i+1} = 1, \\ 0 & \text{else.} \end{cases}$$

Additionally let

$$n_{11.1} = \sum_{i=1}^{c^*-1} I_i^{(11)},\ n_{11.2} = \sum_{i=c^*}^{n-1} I_i^{(11)},$$

where

$$I_i^{(11)} = \begin{cases} 1 & \text{if } \mathrm{J}_i = 1 \ \text{ and } \ J_{i+1} = 1, \\ 0 & \text{else.} \end{cases}$$

Also,

$$n_{1.1} = \sum_{i=1}^{c^*-1} J_i, \quad n_{1.2} = \sum_{i=c^*}^{n-1} J_i,$$

$$n_{0.1} = (c^* - 1) - n_{1.1}, \quad n_{0.2} = (n - c^*) - n_{1.2}.$$

Thus,

$$\hat{p}_{0.1} = \frac{n_{01.1}}{n_{0.1}}, \quad \hat{p}_{0.2} = \frac{n_{01.2}}{n_{0.2}},$$

$$\hat{p}_{1.1} = \frac{n_{11.1}}{n_{1.1}}, \quad \hat{p}_{1.2} = \frac{n_{11.2}}{n_{1.2}}.$$

Now, that we have found the proper MLEs, we can write the expression for the likelihood ratio statistic.

$$\begin{aligned} \log(\Lambda_{c^*}) \quad = \quad & n_{01} \log \hat{p}_0 + (n_0 - n_{01}) \log (1 - \hat{p}_0) + n_{11} \log \hat{p}_1 + (n_1 - n_{11}) \log (1 - \hat{p}_1) \\ & - n_{01.1} \log \hat{p}_{0.1} - (n_{0.1} - n_{01.1}) \log (1 - \hat{p}_{0.1}) - n_{11.1} \log \hat{p}_{1.1} \\ & - (n_{1.1} - n_{11.1}) \log (1 - \hat{p}_{1.1}) - n_{01.2} \log \hat{p}_{0.2} - (n_{0.2} - n_{01.2}) \log (1 - \hat{p}_{0.2}) \\ & - n_{11.2} \log \hat{p}_{1.2} - (n_{1.2} - n_{11.2}) \log (1 - \hat{p}_{1.2}). \end{aligned}$$

Once again, when testing for an unknown change-point,

$$-2\log(\Lambda) = \max_{l \le c \le h} \{-2\log(\Lambda_c)\}.$$

It is especially prudent to pick a sufficiently large $l$ and $n-h$ becuase if $n_{01.1} = 0, n_{11.1} = 0$, etc. for any $c$, it is not possible to calculate $T = -2\log(\Lambda)$. Because the null hypothesis has exactly two free parameters, if we want to find the P-value of the test, we find

$$\xi^{(2)}_{[l/n,h/n]}(T).$$

For instance, the 95% cut-off is the value $x^*$ which solves $\xi^{(2)}_{[l/n,h/n]}(x^*) = .05$. Using $l/n = 1 - h/n = .1$, we can conclude that there is a change-point in the Markov chain structure if $-2\log(\Lambda) > 12.42093$.

If we want to test for a change-point in a general Type B precipitation model, where $\{J_i\}_{i=1}^n$ follows the previously described Markov chain structure and where $\{Z_i\}_{i=1}^n$ is $iid$ with distribution fucntion $G(x|\theta)$, we would apply the following test statistic:

$$\begin{aligned}
\log(\Lambda_{c^*}) \quad = \quad & n_{01}\log\hat{p}_0 + (n_0 - n_{01})\log(1-\hat{p}_0) + n_{11}\log\hat{p}_1 + (n_1 - n_{11})\log(1-\hat{p}_1) \\
& -n_{01.1}\log\hat{p}_{0.1} - (n_{0.1} - n_{01.1})\log(1-\hat{p}_{0.1}) - n_{11.1}\log\hat{p}_{1.1} \\
& -(n_{1.1} - n_{11.1})\log(1-\hat{p}_{1.1}) - n_{01.2}\log\hat{p}_{0.2} - (n_{0.2} - n_{01.2})\log(1-\hat{p}_{0.2}) \\
& -n_{11.2}\log\hat{p}_{1.2} - (n_{1.2} - n_{11.2})\log(1-\hat{p}_{1.2}) + L(\mathbf{y}|\hat{\theta}) \\
& -L(\mathbf{y}_1|\hat{\theta}_1) - L(\mathbf{y}_2|\hat{\theta}_2).
\end{aligned}$$

If we were to test for a change-point in data set while assuming a Type B precipitation

model with $G(x) \sim Gamma(\alpha, \beta)$, then there would be a total of four free parameters under the null hypothesis. Because $\xi^{(4)}_{[.1,.9]}(17.15037) = .05$, using $l/n = 1 - h/n = .1$, we would reject the null hypothesis if $-2\log(\Lambda) > 17.15037$.

# Chapter 9

# Results and Conclusions

For the data set, $n = 6353$ and $n_r = 2823$. If we assume an $MExp$ model, we find $\hat{p} = .444357$ and $\hat{\lambda} = 10.14956$.

## 9.1 Application of the Change-Point Tests

Using a $MExp$ model and discarding 10% of the data on each end, we find $-2\log(\Lambda) = 66.57622$ with $\hat{c} = 4409$. Because $66.57 >> 12.42$, which is the 95% cut-off using $d = 2$, we can reject $H_0$ with a very high degree of confidence. The change-point test finds that $\hat{p}_1 = .4767521$, $\hat{p}_2 = .3708848$, $\hat{\lambda}_1 = 9.92352$ and $\hat{\lambda}_2 = 10.87153$. Figure 9.1 is a graph of $c^*$ vs. $-2\log(\Lambda_{c^*})$ under $MExp$ assumptions.

However, do the results change if we instead assume that the series follows a $MGam$ model? When 10% is ignored on each end $-2\log(\Lambda) = 66.79801$ where $\hat{c} = 4409$. The 95% rejection value for this model is $x^* = 14.92045$, hence the null hypothesis still is soundly rejected. Also, the estimates are $\hat{p}_1 = .4767521$ and $\hat{p}_2 = .3708848$, $\hat{\gamma}_1 = .7927821$ and $\hat{\gamma}_2 = .837829$, and $\hat{\beta}_1 = .1271102$ and $\hat{\beta}_2 = .1160260$. The graph of $c^*$ against $-2\log(\Lambda_{c^*})$ under $MGam$ assumptions is shown in Figure 9.2.

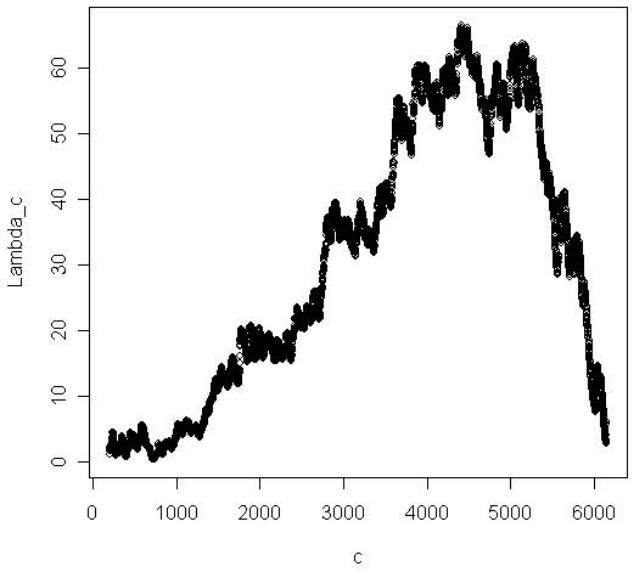


Figure 9.1: $c^*$ vs. $-2\log(\Lambda_{c^*})$ for $MExp$ Model

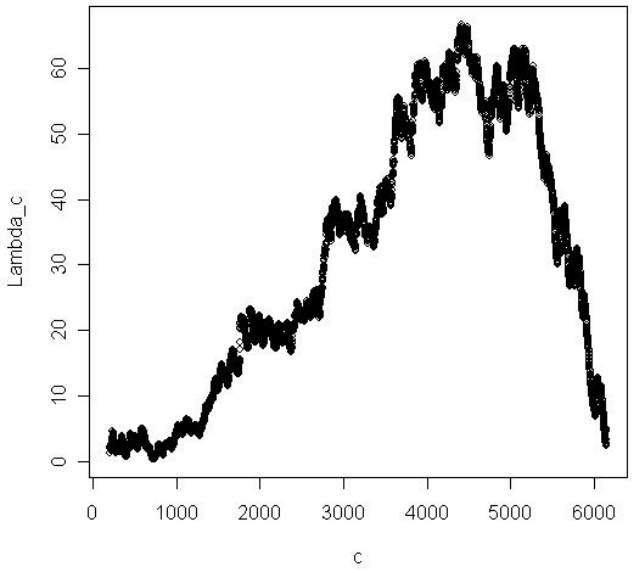


Figure 9.2: $c^*$ vs. $-2\log(\Lambda_{c^*})$ for $MGam$ Model

The results of the change-point test are nearly identical for the $MExp$ and $MGam$ models of precipitation. Will this always be the case, or do the models produce similiar results only because the drastic change in $p$ nullifies the relatively small change in the amount of rain on wet days?

If we examine $\{Y_i\}_{i=1}^{2823}$, which is the series of rainfall magnitudes on rainy days, we can assume that it is a sample from an $exp(\lambda)$ distribution and then test for a change in $\lambda$. This is the same as testing the alternative hypothesis that $p_1 = p_2$ and $\lambda_1 \neq \lambda_2$ for some $c$ under $MExp$ assumptions. If 10% is ignored on each end, then $-2\log(\Lambda) = 15.34574$ where $\hat{c} = 1856$ (which corresponds to the $3647^{th}$ day overall). Since the P-Value $\approx \xi^{(1)}_{[.1,.9]}(15.34574) = .003177$, the test shows a significant change in the amount of rain. The test shows $\hat{\lambda}_1 = 9.6491$ and $\hat{\lambda}_2 = 11.2717$. The graph of $c^*$ against $-2\log(\Lambda_{c^*})$ for this test is shown in Figure 9.3.

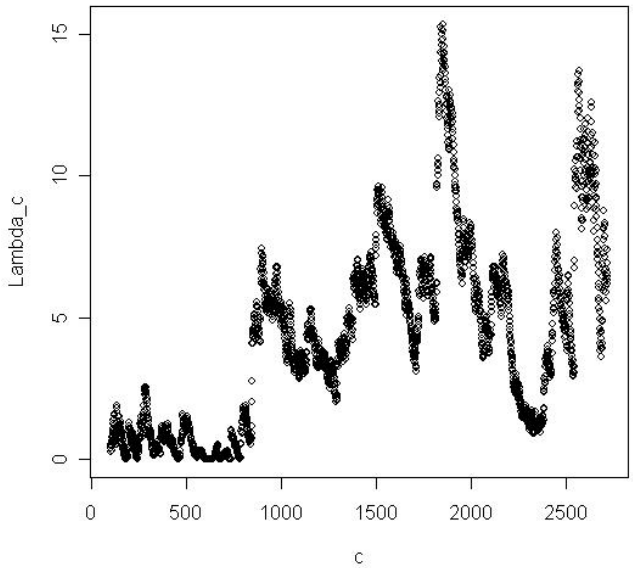


Figure 9.3: $c^*$ vs. $-2\log(\Lambda_{c^*})$ for $exp(\lambda)$ Model of $Y_i$

If we assume that $\{Y_i\}_{i=1}^{2823}$ is a sample from a $gamma(\gamma, \beta)$ distribution, we can test for

a change-point in $\gamma$ or $\beta$. If 10% is ignored on each end, then $-2\log(\Lambda) = 15.20998$ where $\hat{c} = 1856$. The P-Value $\approx \xi^{(2)}_{[.1,.9]}(15.20998) = .015449$, hence the test still shows a significant change in the amount of rain on rainy days. The test finds $\hat{\gamma}_1 = .7838523$, $\hat{\gamma}_2 = .8513023$, and $\hat{\beta}_1 = .1323837$ and $\hat{\beta}_2 = .1042141$. The graph of $c^*$ against $-2\log(\Lambda_{c^*})$ for this test is shown in Figure 9.4.

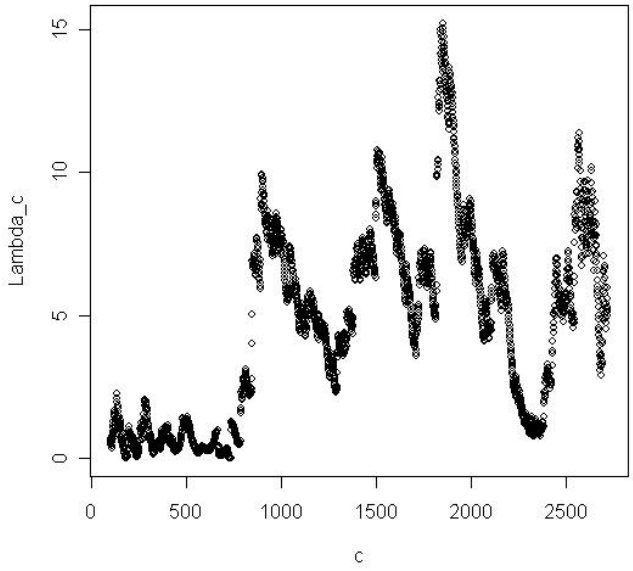


Figure 9.4: $c^*$ vs. $-2\log(\Lambda_{c^*})$ for $Gamma(\gamma, \beta)$ Model of $Y_i$

As you can see, nearly identical change-point are obtained from gamma or exponential assumptions in this series of preciptation data.

We can also test for a change-point in only the occurence of rain, which is to test under an alternative of $p_1 \neq p_2$ where all other parameters do not change. This test produces $-2\log(\Lambda) = 62.17394$ where $\hat{c} = 4409$, which is a highly significant change-point. Once again, $\hat{p}_1 = .4767521$ and $\hat{p}_2 = .3708848$. The graph of $c^*$ against $-2\log(\Lambda_{c^*})$ for this test is shown in Figure 9.5.

If we assume a Type B model, and we test for a change-point in the Markov structure,

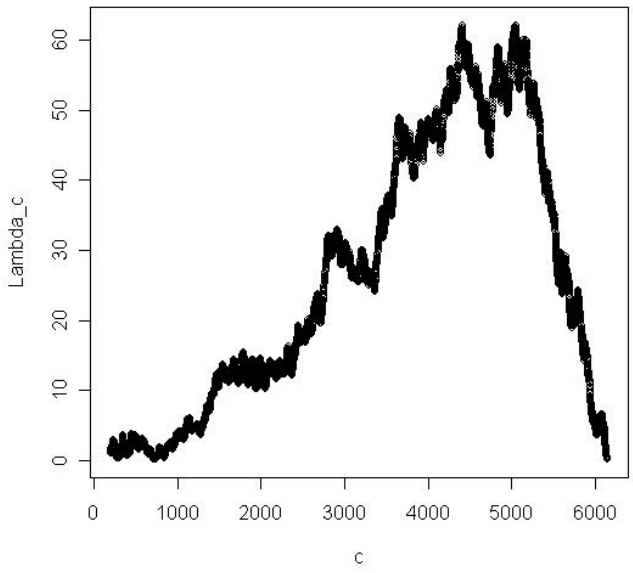


Figure 9.5: $c^*$ vs. $-2\log(\Lambda_{c^*})$ for testing just $p_1 \neq p_2$

then using $l/n = 1 - h/n = .1$, we find $-2\log(\Lambda) = 56.31739$ with $\hat{c} = 5051$. This test concludes that there is a rather significant change-point. A graph of $c$ vs. $-2\log(\Lambda_c)$ in this test is shown in Figure 9.6. The graph shows that despite the different change-point, the actual $\Lambda_c$ series hasn't changed much from when a Bernoulli stucture was assumed. Also, the test finds $p_{0.1} = .4487$, $p_{1.1} = .4930$, $p_{0.2} = .3278$, and $p_{1.2} = .3877$.

If we assume a Class B models with gamma distributed intensities, then for $l/n = 1 - h/n = .1$, we see $-2\log(\Lambda) = 60.68909 > 17.15037$ with $\hat{c} = 4409$. The test finds $p_{0.1} = .4530$, $p_{1.1} = .5029$, $p_{0.2} = .3549$, and $p_{1.2} = .3981$. Figure 9.7 shows $c$ vs. $-2\log(\Lambda_c)$ for this test.

An interesting observation: If the intensities are assumed to have a gamma distribution, the test for a change in the intensities found $\hat{c} = 3647$. The test for a change in the rain occurrence under Bernoulli assumptions found $\hat{c} = 4409$, which is also the result for the test of an overall change in the Type A model. However, the test for a change in the occurrence

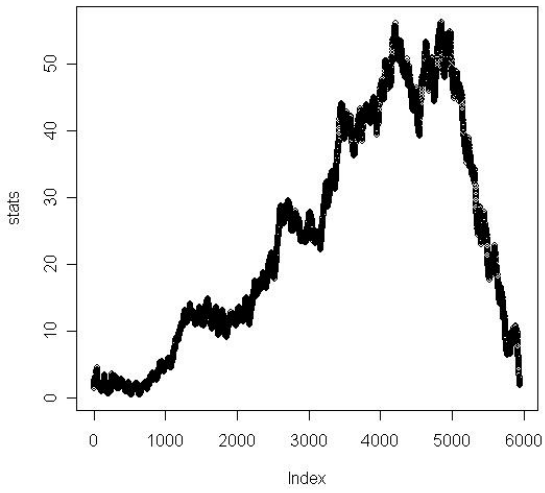


Figure 9.6: $c^*$ vs. $-2\log(\Lambda_c)$ when Testing for Change in Markov Structure

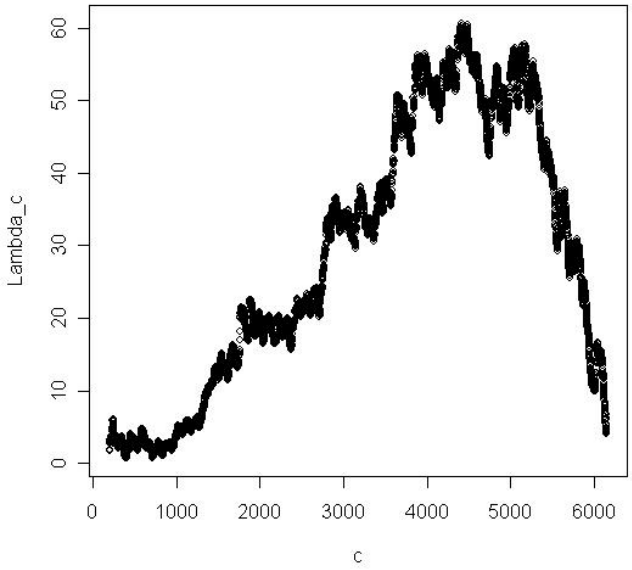


Figure 9.7: $c^*$ vs. $-2\log(\Lambda_{c^*})$ for Testing Change in Type B Model

of rain under Markov assumptions indicated $\hat{c} = 5051$. Oddly enough, the test for an overall change in the Type B model again yielded $\hat{c} = 4409$.

## 9.2 Conclusions

The test showed that there was a significant change in the amount of rain on wet days in our data. However, there is a far more drastic change in the probability of rain that drives the overall behavior of the test statistic. This is one of the advantages of the change-point test as developed for the overall model: it produces a single change-point when testing for change-points in the individual parameters might produce multiple change-points.

If it is plausible to consider that each parameter might change at separate times, it might be best to test for changes in each parameter separately. However, it may be difficult to imagine an instance where naturally the probably of rain would change at one time, and then the amount that it rains on wet days would change at another.

It may seem excessively arduous to test for an over-all change-point while also performing other tests for changes in individual parameters, but fortunately, those and other tests are able to be performed with relative ease using the methods developed in this paper. This is because the Brownian bridge based asymptotic distribution of $-2\log(\Lambda)$ that produces quantiles via the function, $\xi^{(d)}_{[l/n,h/n]}(x^*)$ provides a very versatile test that allows us to test assumptions for a myriad of models.

There are, of course, several drawbacks. Since it is an asymptotic test, a large sample size is required. Even when there is a large sample size, the exact Type I error probability is still unknown and cannot be precisely estimated. Somewhat disturbing is the fact that under the null hypothesis, $\hat{c}$ has a blatantly non-uniform distribution. Also, the test becomes highly volatile for change-points that may occurring near one of the end-points. However, any change-point test will have poor power if a change occurs near an end-point, so it may not be particularly advantageous to test for such changes. We recommend that 10% of the data on each end-point be ignored, due to the fact that the test appears to be stable

when $l/n = 1 - h/n = .1$ and you still can search for change over a wide range of the data. However, the method described in this paper will allow for one to perform a test over any range of the data that the experimenter wishes to use.

The main advantage of using $\xi^{(d)}_{[l/n,h/n]}(x^*)$ to produce test quantiles is that almost any model can be tested for change in this fashion. All that is necessary is that one have a method for producing maximum likelihood estimates for the parameters in the assumed model. Then one can promptly calculate $\Lambda_c$ which promptly begets $\Lambda$. In this paper, methods for testing $MExp$ and $MGam$ models are explained; however if one wishes to assume that rain behaves via a Markov chain, one need only determine the method for finding maximum likelihood estimates for the two probability parameters in the two-state Markov model. If a gamma distribution is assumed for $Z_i$, then quantiles can be produced using $\xi^{(4)}_{[l/n,h/n]}(x^*)$.

Statistics is, by definition, an imprecise science, and it is used to analyze precise sciences, such as climatology. Even though the climate behaves in a definite fashion, no statistical test to determine that behavior can ever be equally as definite. Even though the methods presented in this paper may not be perfect, they do apply some of the most sophisticated and powerful techniques in the realm of modern statistical science.

# Bibliography

Alexandersson, Hans. 1986. A homogeneity test applied to precipitation data. *Journal of climatology*, **6**, 667–675.

Anderson, T.W., & Goodman, Leo A. 1957. Statsistical inference about markov chains. *The annals of mathematical statistics*, **28**, 89–110.

Burgueno, A. Serra, C., & Lana, X. 2004. Monthly and annual statistical distributions of daily rainfall at the fabra observatory (barcelona, ne spain) for the years 1917-1999. *Theoretical and applied climatology*, **77**, 57–75.

Casella, George, & Berger, Roger L. 2002. *Statistical inference*. Duxbury.

Chapman, Tom G. 1997. Stochastic models for daily rainfall in the western pacific. *Mathematics and computers in simulation*, **43**, 351–358.

Csorgo, Miklos, & Horvath, Lajos. 1997. *Limit theorems in change-point analysis*. John Wiley  Sons Ltd.

Gregory, J.M. Wigley, T.M.L., & Jones, P.D. 1992. Determining and interpreting the order of a two-state markov chain: Application to models of daily precipitation. *Water resources research*, **28**, 1443–1446.

Harmel, R.D. et al. 2000. Hydrological response of a small watershed model to generated precipitaion. *Transactions of the american society of argicultural engineers*, **43**, 1483–1488.

Katz, Richard W. 1977. Precipitation as a chain-dependent process. *Journal of applied meteorology*, **16**, 671–676.

Katz, R.W. 1999. Extreme value theory for precipitation: Sensitivity analysis for climate change. *Advances in water resources*, **23**, 133–139.

Mielke, Paul, & Johnson, Earl, S. 1973. Three-parameter kappa distribution maximum likelihood estimates and likelihood ratio tests. *Monthly weather review*, **101**, 701–707.

Srikanthan, R., & McMahaon, T.A. 2001. Stochastic generation of annual, monthly, and daily climate data: A review. *Hydrology and earth system sciences*, **5**, 653–670.

Stern, R.D., & Coe, R. 1984. A model fitting analysis of daily rainfall data. *Journal of the royal statistical society*, **147**, 1–34.

Thom, H.C.S. 1968. Approximate convolution of the gamma and mixed gamma distributions. *Monthly weather review*, **96**, 883–886.

Wilks, Daniel S. 2006. *Statistical methods in the atmospheric sciences*. Elsevier Academic Press.

Wilks, D.S. 1990. Maximum likelihood estimation for the gamma distribution using data containing zeros. *Journal of climate*, **3**, 1495–1501.

Wilks, D.S. 1998. Multisite generalization of a daily stochastic precipitation generation model. *Journal of hydrology*, **210**, 178–191.

Wilks, D.S. 1999. Interannual variability and extreme-value characteristics of several stochastic daily precipitaion models. *Agricultural and forest meteorology*, **93**, 153–169.

Woolhiser, David A., & Pegram, G.G.S. 1979. Maximum likelihood estimation of fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. *Journal of applied meteorology*, **18**, 34–42.