UCONN LIBRARY

University of Connecticut
OpenCommons@UConn

Doctoral Dissertations | University of Connecticut Graduate School

8-8-2019

# Scan Statistics for Detecting a Local Change in the Scale Parameter for Gamma Random Variables

Qian Meng
*University of Connecticut - Storrs*, qian.meng@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
Meng, Qian, "Scan Statistics for Detecting a Local Change in the Scale Parameter for Gamma Random Variables" (2019). *Doctoral Dissertations*. 2243.
https://opencommons.uconn.edu/dissertations/2243

# Scan Statistics for Detecting a Local Change in the Scale Parameter for Gamma Random Variables

Qian Meng, Ph.D.
University of Connecticut, 2019

## ABSTRACT

In this dissertation scan statistics for detecting a local change in the scale parameter for gamma and exponential random variables are investigated for both one dimensional and two dimensional cases. The shape parameter is assumed to be known for the gamma random variables. When the size of the window where the local change occurs is known but the scale parameter in null hypothesis is unknown, conditional fixed window scan statistics are proposed. When the true window size is unknown, conditional multiple window minimum $p$-value scan statistics and variable window scan statistics based on generalized likelihood ratio test principle are developed. The performance of the proposed scan statistics is evaluated by Monte Carlo simulation studies. For moderate to large shift in scale parameter, conditional fixed window scan statistics with the correct scanning window size, multiple window and variable window scan statistics all performed well.

# Scan Statistics for Detecting a Local Change in the Scale Parameter for Gamma Random Variables

**Qian Meng**

B.S., Peking University, Beijing, China, 2012

M.S., University of Connecticut, CT, USA, 2014

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2019

Copyright by

Qian Meng

2019

# APPROVAL PAGE

Doctor of Philosophy Dissertation

# Scan Statistics for Detecting a Local Change in the Scale Parameter for Gamma Random Variables

Presented by

Qian Meng, B.S., M.S.

Major Advisor _____

Joseph Glaz

Associate Advisor _____

Nitis Mukhopadyay

Associate Advisor _____

Richard A. Vitale

University of Connecticut

2019

# Acknowledgements

I would like to express my deepest gratitude and appreciation to my advisor, Professor Joseph Glaz, for all his guidance, advising and encouragement. Without his patient and warm support, I would not be able to accomplish my PhD study. What I learned from him is not only knowledge, but also attitude to research and to life. He guided me through the dark times, helped me understand what is critical thinking, and he is no doubt one of the most important people in my life journey.

I must also deeply acknowledge Professor Nitis Mukhopadhyay and Professor Richard Vitale, for being my associate advisors on my dissertation committee, and for all their help and constructive advice in improving my dissertation. Appreciation is due to Professor Jun Yan, who provided research assistantship for me in 2014. I am extremely grateful to Professor Zhiyi Chi, and all other professors and staff in the Department of Statistics, for building an excellent doctoral program for both studying and research. In addition, I obtained valuable teaching experience here, which profoundly helped me become my better self.

Last but not least, I would like to convey my wholehearted appreciation to my parents, Ms. Minlan Huang and Mr. Yong Meng, for all their unconditional and never-changed love. They have given me so much that I have benefited from, and will keep benefiting from in the future. I would also like to acknowledge my boyfriend, Joseph

Quigley V, for all his love and support. I am so lucky to have you beside me! Sincere thanks should also go to Yishu Xue, Ruochen Zha, Cuicui Wang and all my other friends – for being supportive, and for making this part of my life so much better. Special thanks to fluffy Lady Melitta and Sir Franz, for their lovely companionship, and never being absent for my up and down moments.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Research on moving sums and scan statistics has been of great interest in scientific literature in applied probability and statistics in the past three decades. The goal is to detect the occurrence of a local change in parameters of a statistical model in a subsequence of all observations. There have been many applications in different areas of science and technology, including but not limited to epidemiology, astronomy, genetics, material science, linguistics, quality control, telecommunication, geology (Hoh and Ott, 2000; Glaz et al., 2009; Glaz and Balakrishnan, 2012; Kulldorff, 1997).

Assuming all observations follow the same distribution, the essence of scan statistics is to use a sliding window that contains multiple observations, to scan through the whole area, and decide whether there is a sub-region where a change of distribution parameters happens. The location of the local change is usually unknown and needs to be detected. When the window size of the local change in the parameter is known, a fixed scanning window can be employed for the scanning procedure, which is often referred to as the fixed window scan statistic approach. However, in many practical scenarios, the true window size where the local change in parameter occurs is unknown.

To avoid loss of power caused by the use of incorrect scanning window size, and the complexity of multiple testing, it is necessary to incorporate multiple window sizes in the scanning process. There are two different approaches that have been implemented – multiple window scan statistics (Glaz and Zhang, 2004) and variable window scan statistics (Nagawalla, 1996).

Extensive research has been done on moving sums and scan statistics in one dimensional case, including Glaz and Naus (1991), Haiman (2007), and Glaz and Balakrishnan (2012). To address emerging problems in the area of scan statistics, new methodologies have been developed for data in two dimensional regions, including probability approximations and inequalities (Chen and Glaz, 1996; Haiman and Preda, 2002, 2006). Different probability models for one and two dimensional data have been studied, including both discrete and continuous distributions. Scan statistics for Bernoulli, binomial, Poisson and normal random variables have been investigated for both one and two dimensional data (Chen and Glaz, 1996; Glaz et al., 2012; Zhao and Glaz, 2017). Scan statistics for detecting a local change in mean and variance for normal random variables have been studied in Wang and Glaz (2014), Zhao and Glaz (2016) and Zhao and Glaz (2017). The use of scan statistics for detecting a local change in the parameters for exponential or gamma observations have not been studied in the statistical literature. There is a wide range of potential applications of scan statistics for data modeled by exponential or gamma distributions in many areas, including: genomics, environmental science, finance and quality control (Mendoza-Parra et al., 2013; Aksoy, 2000; Boland,

2007). Poisson process, generated by the exponential distribution, also has potential applications for scan statistics (Daley and Vere-Jones, 2003).

This dissertation's research is focusing on developing scan statistics for detecting a local change in the scale parameter $\beta$ for one and two dimensional observations modeled by a gamma distribution. The shape parameter $\theta$ is assumed to be known and remains constant for all observations, throughout all chapters in this dissertation. Hence, a local increase in the scale parameter is equivalent to a local increase in the population mean. Levin et al. (2005) proposed a model-based scan statistic for gene clustering using simple and compound Poisson processes, revealing that there are potential applications of scan statistics for exponential and gamma random variables in the field of genetics. Joint distributions conditional on the sufficient statistic for the unknown parameter, for the binomial, negative binomial, Poisson and normal models, have been employed in Chen and Glaz (2016); Zhao and Glaz (2016); Chen and Glaz (2017), to develop scan statistics for detecting a local change in respective parameters. Hoh and Ott (2000) provided an interesting approach in using moving sums to detect susceptibility genes, where a minimum $p$-value statistic is utilized to combine information on multiple contiguous genetic markers. More complete and organized methodologies on multiple window scan statistic for normal data was presented in Zhao and Glaz (2016). In addition, Nagawalla (1996) and Kulldorff (1997) provide motivation and inspiration on adopting generalized likelihood ratio tests to construct variable window scan statistics, in the context of multiple testing problems.

This dissertation is organized as follows: Chapter 2 formally formulates the test of hypothesis of interest for detecting a local change in the scale parameter for a sequence of gamma random variables and the special case of exponential random variables, in the one dimensional case, assuming that the true window size of the local change is known. A conditional fixed window scan statistic is proposed for the case when the scale parameter in null hypothesis is unknown. Chapter 3 investigates multiple window and variable window scan statistics, under the circumstances that the true window size of the local change is unknown. Chapter 4 presents the fixed window, multiple window and variable window scan statistics for gamma random variables in two dimensional case. In Chapters 2, 3 and 4, algorithms to calculate the critical value of the corresponding scan statistics, as well as the powers of the tests. Numerical results based on Monte Carlo simulation studies based on these algorithms are presented and discussed. In Section 3.5, we include an example of the use of the multiple window and variable window scan statistics, for a data set of time intervals between coal mine disasters. In conclusion, Chapter 5 provides a brief summary of the methodologies and results, as well as a discussion on future work related to the results obtained in this dissertation.

# Chapter 2

# Fixed Window Scan Statistics for One Dimensional Gamma Random Variables

## 2.1 Introduction

The purpose of scan statistics is to detect local changes in model parameters for observed data. The focus of this chapter is to investigate the performance of a fixed window scan statistic to detect a local change in the scale parameter for a series of observations that follow a gamma distribution.

Let $X_1, X_2, ..., X_N$ be a sequence of independent and identically distributed (i.i.d.) observations that follow the gamma distribution $\Gamma(\theta, \beta)$, with density function given by:

$$f_{X_i}(x_i) = \frac{1}{\Gamma(\theta)\beta^{\theta}} x_i^{\theta-1} \exp\left(-x_i/\beta\right), \quad i = 1, \ldots, N, \tag{2.1}$$

where $x_i > 0$, $\theta, \beta > 0$ , and $N$ is the predetermined length of the data being monitored. Recall that we assume that the shape parameter $\theta$ is known and constant for all observations. The objective is to detect whether there is a local upward shift of the scale parameter $\beta$, which is equivalent to detecting an increase in population mean. The methods developed in this dissertation can be easily modified to detect a downward or a two-sided shift.

Let $2 \leq m \leq N/4$ be the length of a sliding window for a sequence of $m$ consecutive observations. We are interested in testing the following hypotheses:

$$
\begin{aligned}
&H_0 : X_i \sim \Gamma(\theta, \beta_0), \forall i = 1, 2, ..., N; X_i\text{'s are independent; vs.} \\
&H_a : \exists j, 1 \leq j \leq N - m + 1, \text{such that } X_i \sim \Gamma(\theta, \beta_1), \beta_1 > \beta_0, \forall i = j, ..., j + m - 1; \\
&\qquad \text{and } X_i \sim \Gamma(\theta, \beta_0), \forall i = 1, ..., j - 1, j + m, ..., N; X_i\text{'s are independent;}
\end{aligned}
\tag{2.2}
$$

where $\theta$ is known, and $\beta_1$ is unknown. We will discuss both situations where $\beta_0$ is known and $\beta_0$ is unknown.

The goal is to determine whether there is a window of length $m$, with an unknown starting position $j$, where the observations inside the window follow a gamma distribution with a larger scale parameter, which would lead to larger mean and variance, than those outside of the window. In this chapter we assume that $m$ is known. The use of constraint $m \leq N/4$ is to emphasize that the focus is on detecting a local change in scale

parameter in a window of small or moderate length.

The shape parameter $\theta$ is assumed to be known and constant for observations inside and outside the window, while a possible change in the scale parameter has occurred. However, in most applications, the scale parameter $\beta_0$, specified in the null hypothesis, is unknown. This difficulty can be resolved by conditioning on the sum of observed data. Under $H_0$, the joint distribution function of $X_1, ..., X_N$, conditioned on their sum, does not depend on any unknown parameter.

The rest of this chapter discusses fixed window scan statistics for gamma and exponential random variables for testing hypotheses formulated above in (2.2), and is organized as follows. In Section 2.2, a fixed window scan statistic is proposed for a sequence of gamma random variables, based on their joint distribution, conditioned on the sum of all observations. The sampling algorithm to evaluate the rejection region for the hypothesis testing problem in Equation (2.2) is presented in Section 2.2.1. In Section 2.3, the conditional fixed window scan statistic for a sequence of exponential random variables, is derived on the basis of the methodologies in Section 2.2. A simulation study to evaluate the performance of the proposed fixed window statistics is presented in Section 2.4. Concluding remarks are discussed in Section 2.5.

## 2.2 Gamma Random Variables with Known Shape Parameter

Let $2 \leq m \leq N/4$ be the length of the sliding window for a sequence of gamma random variables $X_1, X_2, \ldots, X_N$. In this section we assume that the length of the window $m$, where a potential change in the scale parameter has occurred, is known. To test the hypotheses in (2.2), we first define the sequence of moving sums:

$$Y_{j,m} = \sum_{i=j}^{j+m-1} X_i, 1 \leq j \leq N - m + 1, \tag{2.3}$$

where $j$ denotes an unknown starting position of a window of size $m$, where a potential shift in the scale parameter has occurred, in the sequence of the observed data. The summation of all the observed data is denoted by:

$$Y = \sum_{i=1}^{N} X_i. \tag{2.4}$$

If the scale parameter under the null hypothesis is known, then the following *unconditional fixed window scan statistic* could be used for detecting a local change in scale parameter:

$$S_{m,N} = \max\{Y_{j,m}; 1 \leq j \leq N - m + 1\}. \tag{2.5}$$

Under $H_0$ the $X_i$'s are i.i.d. gamma random variables. The sequence of moving sums $\{Y_{j,m}, 1 \leq j \leq N-m+1\}$ is stationary and $m$-dependent and it has a special joint multivariate gamma distribution, with identical marginal distributions $\Gamma(m\theta, \beta_0)$.

For $2 \leq m \leq N/4$ and $-\infty < t < +\infty$, let

$$G_{m,t}(N) = P(S_{m,N} < t) = P(Y_{1,m} < t, Y_{2,m} < t, ..., Y_{N-m+1,m} < t), \tag{2.6}$$

denote the cumulative distribution function for $S_{m,N}$. The probability that the unconditional fixed window scan statistic exceeds level $t$ is given by:

$$P(S_{m,N} \geq t) = 1 - G_{m,t}(N). \tag{2.7}$$

When the values of $m, M$ and $t$ are clearly understood, to simplify the notations, we abbreviate $G_{m,t}(N)$ as $G(N)$, and $S_{m,N}$ as $S_m$.

In our hypothesis testing problem, under $H_0$ the scale parameter $\beta_0$ is unknown. Therefore, we cannot evaluate the tail probability given in Equation (2.7). In this case, we propose to condition on the sufficient statistic for $\beta$, which is the sum of all observations, $Y$, to eliminate the unknown parameter for the null distribution of a *conditional* scan statistic defined below. The use of conditional tests when the distribution of the test statistic under the null hypothesis has unknown parameters is well documented in statistical literature. The most direct approach is to condition on the sufficient statistic

under $H_0$ (Davison and Hinkley, 1997, p.138). In the scan statistics literature conditional tests have also been adopted among others (Chen and Glaz, 2016). We now proceed to define the conditional scan statistic for the problem at hand.

Under $H_0$, $Y \sim \Gamma(N\theta, \beta_0)$, and the joint distribution of $X_1, X_2, ..., X_N$ conditional on $Y = y$ is given by:

$$\begin{aligned} f_{X_1,...,X_N|Y=y}(x_1, x_2, ..., x_N) &= f_{X_1,...,X_N,Y}(x_1, x_2, ..., x_N, y)/f_Y(y) \\ &= f_{X_1,...,X_N}(x_1, ..., x_{N-1}, y - \sum_{i=1}^{N} x_i)/f_Y(y) \\ &= \prod_{i=1}^{N-1} f_{X_i}(x_i) \cdot f_{X_N}(y - \sum_{i=1}^{N-1} x_i)/f_Y(y). \end{aligned}$$

This can be simplified to:

$$f_{X_1,X_2,...,X_N|Y=y}(x_1, x_2, ..., x_N) = \frac{\Gamma(N\theta)}{[\Gamma(\theta)]^N} \cdot \frac{\left[\prod_{i=1}^{N-1} x_i \cdot (y - \sum_{i=1}^{N-1} x_i)\right]^{\theta-1}}{y^{N\theta-1}}, \tag{2.8}$$

where $0 < x_i < y,\ \forall i = 1, ..., N$, and $\sum_{i=1}^{N} x_i = y$. Therefore, the conditional distribution in (2.8) is now free of the scale parameter $\beta_0$. Similarly, under the null hypothesis, the joint distribution of $X_j, ..., X_{j+m-1},\ 1 \leq j \leq N - m + 1$, conditional on the partial

sum $Y_{j,m}$ does not depend on $\beta_0$ and is given by:

$$f_{X_j,X_{j+1},...,X_{j+m-1}|Y_{j,m}=y_{j,m}}(x_j, x_{j+1}, ..., x_{j+m-1}) = \frac{\Gamma(m\theta)}{[\Gamma(\theta)]^m} \cdot \frac{\left[\prod_{i=j}^{j+m-1} x_i \cdot (y_{j,m} - \sum_{i=j}^{j+m-1} x_i)\right]^{\theta-1}}{y_{j,m}^{m\theta-1}}.$$

Let $Z_i = X_i/Y$, then $\sum_{i=1}^{N} Z_i = 1$. It is well known that the joint distribution of $Z_1, Z_2, \ldots, Z_N$ is $\text{Dir}(\vec{\theta}_N)$, where $\vec{\theta}_N = (\theta, ..., \theta)^N$ is an N dimensional vector, $\theta$ being the shape parameter of the gamma distribution of the $X_i$'s (Devroye, 2013, p. 593).

Consequently, we can define the *conditional fixed window scan statistic*:

$$S_m^* = \max\{Y_{j,m}^*; 1 \leq j \leq N - m + 1\}, \tag{2.9}$$

where

$$Y_{j,m}^* = \sum_{i=j}^{j+m-1} Z_i, 1 \leq j \leq N - m + 1. \tag{2.10}$$

We propose to use this scan statistic to test the hypothesis stated in (2.2). The cumulative distribution of $S_m^*$ is given by:

$$G^*(N) = P(S_m^* < t) = P(Y_{1,m}^* < t, Y_{2,m}^* < t, ..., Y_{N-m+1,m}^* < t). \tag{2.11}$$

If the window size $m$ is known, $H_0$ is rejected when $S_m^*$ exceeds a threshold value $t$, where $t$ is determined from $P(S_m^* \geq t) = \alpha$, with $\alpha$ being the specified significance level. For a

given significance level $\alpha$, the critical value denoted by $p_\alpha$, demonstrates the association between the significance level and the rejection region, i.e.

$$P(S_m^* \geq p_\alpha) = \alpha. \tag{2.12}$$

To implement the testing procedure based on $S_m^*$, one needs to evaluate $G^*(N)$.

Since there are no explicit formula or accurate approximations to evaluate $G^*(N)$, one has to employ a Monte Carlo simulation approach to implement the testing procedure based on the scan statistic $S_m^*$.

To investigate the performance of the scan statistic $S_m^*$, we evaluate its power for selected parameters in the alternative hypothesis, via simulation. In the following subsection we discuss the simulation algorithms for determining the rejection region and evaluating the power of this scan statistic.

### 2.2.1 The Simulation Algorithm

To implement the hypothesis testing problem stated in (2.2), based on the conditional fixed window scan statistic $S_m^*$ , the rejection region for a predetermined significance level $\alpha$ has to be determined. The following Algorithm 1 can be used to find the simulated cumulative distribution of $S_m^*$ and the critical value $p_\alpha$, as defined in Equation (2.12).

**Algorithm 1:** Fixed Window Scan Statistics: Critical Value for $S_m^*$

**Result:** Obtain the critical value $S_m^*$ for a given significance level $\alpha$

**for** $r \leftarrow 1$, $R$ **do**

  Draw a sample of $Z_i$'s from Dir($\vec{\theta}_N$) distribution, $i = 1, ..., N$;

  Calculate $N - m + 1$ moving sums $Y_{j,m}^*$;

  Find the maximum moving sum $S_m^{(r)*}$, and sort them in increasing order as a vector $U_m^{(r)}$, thus the simulated cumulative probability distribution (2.13) is based on $U_m^{(r)}$;

**end**

The critical values $p_\alpha$ can be calculated from the simulated $100(1-\alpha)$th percentile of $U_m^{(r)}$.

**Algorithm 2:** Fixed Window Scan Statistics: Power of Test

**Result:** Power of the test for alternative hypothesis set by $\beta_1/\beta_0 = 1, 2, 3, 4, 5$

Choose an arbitrary value for the starting position $j$ of the local change, $1 \leq j \leq N - m + 1$, and the total number of simulations $L$;

**for** $l \leftarrow 1$, $L$ **do**

  Generate $X_1, ..., X_{j-1} \sim \Gamma(\theta, \beta_0)$, $X_j, ..., X_{j+m-1} \sim \Gamma(\theta, \beta_1)$, and $X_{j+m}, ..., X_N \sim \Gamma(\theta, \beta_0)$;

  Calculate $Y = \sum_{i=1}^{N} X_i$, and the maximum moving sum $S_m^{(l)*}$;

  Compare $S_m^{(l)*}$ with the simulated cumulative distribution $U_m^{(r)}$ we obtained in Algorithm 1, and calculate the $p$-value $p_m^{(l)}$ based on (2.15);

**end**

Calculate the power $\widehat{\eta}_\alpha$ by (2.16).

Based on Algorithm 1 outlined above, the simulated cumulative distribution of $S_m^*$, denoted by $P^*(S_m^* < t)$, is given by:

$$P^*(S_m^* < t) = r/R, \quad U_m^{(r)} \leq t < U_m^{(r+1)}, \tag{2.13}$$

where $r$ is the number of times out of $R$ trials in the simulation that $S_m^*$ is less than $t$.

Algorithm 2 is developed to calculate the power of the test, i.e. the probability of correctly rejecting the null hypothesis under alternative hypothesis with a specific $\beta_1/\beta_0$ shift. The simulation studies presented in Section 2.4 include results with alternative hypotheses of $\beta_1 \in \{1, 2, 3, 4, 5\}$ and null hypothesis of $\beta_0 = 1$, since it has been proven above that the exact value of $\beta_0$ does not matter. By conditioning on the total sum of all observations $Y = \sum_{i=1}^{N} X_i$, we can employ the conditional fixed window scan statistic (2.9), which distribution does not depend on any unknown parameter.

Note that in each iteration $l$ of the simulation Algorithm 2, the maximum moving sum is evaluated via:

$$S_m^{(l)*} = \max\{Y_{j,m}^*; j = 1, ..., N - m + 1\}. \tag{2.14}$$

The $p$-value is calculated by:

$$p_m^{(l)} = P(S_m^{(l)*} > t) = r/R, \quad U_m^{(r)} \leq t < U_m^{(r+1)}, \tag{2.15}$$

where $r$ is the number of times out of $R$ trials in the simulation that $S_m^{(l)*}$ is less than $t$. Power calculation is based on:

$$\widehat{\eta}_\alpha = \frac{\#\{p_m^{(l)} < \alpha, l = 1, ..., L\}}{L}. \tag{2.16}$$

## 2.3 Exponential Random Variables

When the shape parameter $\theta = 1$, the gamma distribution $\Gamma(\theta, \beta)$, is the exponential distribution. Assume now we have a sequence of i.i.d. observations, $X_1, X_2, \ldots, X_N$, following the exponential distribution $\text{Exp}(\beta)$, with probability density function

$$f(x_i) = \beta^{-1} \exp\left(-x_i/\beta\right), \quad i = 1, \ldots, N, \ x_i > 0, \ \beta > 0.$$

We are interested in detecting a potential local change in the scale parameter $\beta$, i.e., to find out if there exists a sub-sequence of $m$ observations that follow an exponential distribution with a larger scale parameter than the rest of the observed data. As in Section 2.2, we focus on a local upward shift in $\beta$. The methods to detect a local downward or two-sided shift can be implemented based on the methods presented in this section.

As in Section 2.2, we assume that the length of the window where a local change in the scale parameter has occurred, $m$, is known. The notation for the total sum of the observed data $Y$ is defined in Equation (2.4), the partial sum $Y_{j,m}$ as in (2.3) and

the *conditional fixed window scan statistic* $S_m^*$ as in Equation 2.9. We are interested in testing the following hypotheses:

$$\begin{aligned} H_0 &: X_i \sim \text{Exp}(\beta_0), \forall i = 1, 2, ..., N; X_i\text{'s are independent;} \quad \text{vs.} \\ H_a &: \exists j, 1 \leq j \leq N - m + 1, \text{s.t. } X_i \sim \text{Exp}(\beta_1), \beta_1 > \beta_0, \forall i = j, ..., j + m - 1; \\ &\quad \text{and } X_i \sim \text{Exp}(\beta_0), \forall i = 1, ...j - 1, j + m, ..., N; X_i\text{'s are independent;} \end{aligned} \tag{2.17}$$

where $\beta_0$ and $\beta_1$ are unknown. Under the null hypothesis, similar to the general gamma distribution case discussed in Section 2.2, we have $Y \sim \Gamma(N, \beta_0)$. It follows from Equation (2.8), that the joint probability density function of $X_1, ..., X_N$ conditional on $Y = y$ is given by:

$$f_{X_1,...,X_N|Y=y}(x_1, x_2, ..., x_N|y) = \Gamma(N)/y^{N-1}, \tag{2.18}$$

where $0 < x_i < y,\ \forall i = 1, ..., N$, and $\sum_{i=1}^{N} x_i = y$. Again, this conditional distribution is now free of $\beta_0$. Similarly, under null hypothesis, the joint distribution of $X_j, ..., X_{j+m-1}$ conditional on $Y_{j,m}$ is also free of $\beta_0$. Therefore, on the perspective of conditional probability distribution, no parameter needs to be assumed known. As in the gamma distribution case, we define $Z_i = X_i/Y,\ \forall i = 1, ..., N$, then $Z_1, Z_2, ..., Z_N \sim \text{Dir}(\mathbf{1}_N)$, where $\mathbf{1}_N = (1, ..., 1)^N$ is an $N$ dimensional vector. Therefore, the implementation of the conditional fixed window scan statistic $S_m^*$ and evaluating its power for selected parameters in the alternative hypothesis, via Monte Carlo simulation, is obtained by simulations from

the Dirichlet distribution. The algorithms used to calculate the critical value of the scan statistic and its test power are the same as Algorithm 1 and 2, with the exception of setting $\theta = 1$.

## 2.4 Numerical Results

In Table 1, we present numerical results for rejection region thresholds of conditional fixed window scan statistics with a scanning window of length $m = 5, 7, 10, 15, 20, 25$, respectively, for a sequence of $N = 100$ observations from a gamma distribution with shape parameter $\theta = 0.5, 1, 2, 5$ and scale parameter $\beta_0 = 1$. These thresholds are obtained by Monte Carlo simulations, with $M = 10,000$ replicates. For each fixed window length and $\theta$ value, rejection thresholds for different significance levels $\alpha = 0.10, 0.05, 0.01$ are recorded. As expected, the threshold increases as we have a wider scanning window and lower significance level.

The power for the conditional fixed window scan statistic, for selected parameters of the alternative hypothesis, based on a simulation with $M = 10,000$ trials, is presented in Figure 1. The corresponding numerical results are listed in Tables 2, 3 and 4. As stated in the alternative hypothesis in (2.2), we generate gamma random variables with multiple settings of $\beta_0$ and $\beta_1$. The ratios of $\beta_1/\beta_0$ include 1, 2, 3, 4, and 5, with the ratio = 1, included to evaluate the achieved significance level. The local change is set to start at the $j = 31$st observation, with the true window length of $m = 10$. The scanning

**Table 1:** Rejection thresholds for fixed window scan statistics of gamma random variables ($\theta = 0.5, 1, 2, 5$) by Monte Carlo simulations

| $\theta$ | $\alpha$ | Rejection Threshold | | | | | |
|---|---|---|---|---|---|---|---|
| | | $m = 5$ | $m = 7$ | $m = 10$ | $m = 15$ | $m = 20$ | $m = 25$ |
| 0.5 | 0.10 | 0.1773 | 0.2057 | 0.2462 | 0.3075 | 0.3625 | 0.4181 |
| | 0.05 | 0.1914 | 0.2209 | 0.2639 | 0.3257 | 0.3823 | 0.4375 |
| | 0.01 | 0.2220 | 0.2527 | 0.2979 | 0.3630 | 0.4197 | 0.4793 |
| 1 | 0.10 | 0.1339 | 0.1618 | 0.1991 | 0.2576 | 0.3139 | 0.3682 |
| | 0.05 | 0.1429 | 0.1710 | 0.2103 | 0.2699 | 0.3278 | 0.3810 |
| | 0.01 | 0.1617 | 0.1901 | 0.2320 | 0.2941 | 0.3533 | 0.4085 |
| 2 | 0.10 | 0.1058 | 0.1318 | 0.1678 | 0.2254 | 0.2796 | 0.3323 |
| | 0.05 | 0.1112 | 0.1378 | 0.1747 | 0.2338 | 0.2882 | 0.3411 |
| | 0.01 | 0.1229 | 0.1498 | 0.1898 | 0.2502 | 0.3064 | 0.3596 |
| 5 | 0.10 | 0.0830 | 0.1071 | 0.1413 | 0.1965 | 0.2493 | 0.3015 |
| | 0.05 | 0.0860 | 0.1104 | 0.1453 | 0.2013 | 0.2541 | 0.3073 |
| | 0.01 | 0.0926 | 0.1176 | 0.1533 | 0.2111 | 0.2640 | 0.3182 |

window lengths included are $m = 5, 7, 10, 15, 20$, and 25, which are presented in different colors.

The power of the scan statistic for significance levels $\alpha = 0.01, 0.05$ and 0.10 are shown in the three columns of plots from left to right. The four rows of plots show results for gamma random variables with $\theta = 0.5$, 1, 2 and 5 respectively, and $\theta = 1$ is equivalent to exponential random variables. In the Tables 2 - 5 below, numerical results are presented for gamma distribution random variables with shape parameters $\theta = 0.5$, $\theta = 2$, $\theta = 5$ and $\theta = 1$, respectively.

By inspecting the numerical results for $\beta_1/\beta_0 = 1$, we can see that the achieved type I error rates are close to their corresponding nominal values, i.e., the specified significance levels, indicating that the simulation algorithms are accurate. As the ratio

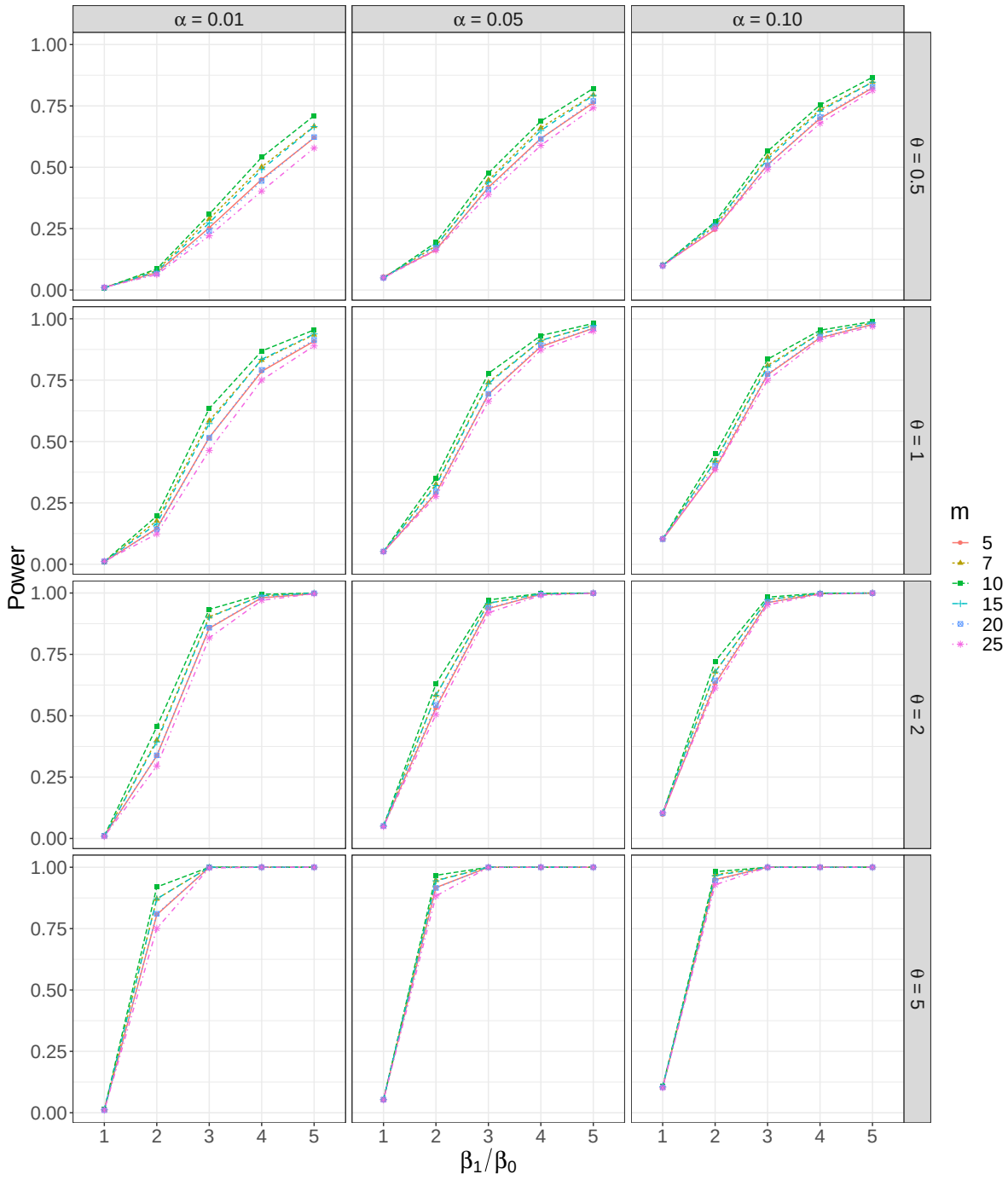


**Figure 1:** Power comparison for fixed window scan statistic (gamma and exponential random variables): $N = 100$, true $m = 10$, window lengths tested $m = 5$, 7, 10, 15, 20 and 25, $\theta = 0.5, 1, 2$ and 5

of $\beta_1/\beta_0$ increases, the probability of successfully detecting the local change increases as well, which is expected since a larger ratio would lead to a relatively bigger shift in local change. We also notice that when the scanning window size is accurate or close to the true window size of the local change, the power is higher. Moreover, when the scanning window size equals the true window size $m = 10$, the power is the highest; and the further away the scanning window size is from the true window size, the lower is its power. For example, when length of the scanning window is $m = 25$, while true window length is $m = 10$, the decrease in power can be as large as 0.10 - 0.15.

In addition, for larger values of the shape parameter $\theta$, the power increases faster as the window length and the $\beta_1/\beta_0$ ratio increase. If we compare across these plots, it can be noticed that with the same $m$, $\beta_1/\beta_0$ and $\alpha$, the power is higher with a larger $\theta$. With $\theta = 2$ and 5, as the $\beta_1/\beta_0$ ratio increases, the power reaches 1 rapidly. One can explain this by considering gamma random variables with an integer $\theta > 1$ as a sum of $\theta$ exponential random variables, indicating that there are more observations. With the same $m$, $\beta_1/\beta_0$ and $\alpha$, $N$ observations following $\Gamma(\theta, \beta_0)$ is equivalent to the $N\theta$ exponential random variables with a local change of size $m\theta$, which would naturally make the detection easier.

**Table 2:** Power for fixed window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\theta = 0.5, L = 10000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.1019 | 0.2462 | 0.5104 | 0.6986 | 0.8235 |
| | $m = 7$ | 0.1016 | 0.2615 | 0.5422 | 0.7349 | 0.8474 |
| | $m = 10$ | 0.0997 | 0.2777 | 0.5652 | 0.7540 | 0.8667 |
| | $m = 15$ | 0.0998 | 0.2706 | 0.5326 | 0.7286 | 0.8470 |
| | $m = 20$ | 0.0991 | 0.2583 | 0.5080 | 0.7042 | 0.8275 |
| | $m = 25$ | 0.0983 | 0.2502 | 0.4910 | 0.6787 | 0.8128 |
| $\alpha = 0.05$ | $m = 5$ | 0.0534 | 0.1634 | 0.4182 | 0.6176 | 0.7629 |
| | $m = 7$ | 0.0505 | 0.1794 | 0.4478 | 0.6611 | 0.7952 |
| | $m = 10$ | 0.0477 | 0.1936 | 0.4764 | 0.6879 | 0.8210 |
| | $m = 15$ | 0.0492 | 0.1824 | 0.4391 | 0.6497 | 0.7927 |
| | $m = 20$ | 0.0495 | 0.1709 | 0.4096 | 0.6153 | 0.7698 |
| | $m = 25$ | 0.0501 | 0.1616 | 0.3889 | 0.5891 | 0.7420 |
| $\alpha = 0.01$ | $m = 5$ | 0.0106 | 0.0706 | 0.2538 | 0.4508 | 0.6198 |
| | $m = 7$ | 0.0104 | 0.0796 | 0.2895 | 0.5019 | 0.6670 |
| | $m = 10$ | 0.0085 | 0.0859 | 0.3087 | 0.5407 | 0.7102 |
| | $m = 15$ | 0.0091 | 0.0762 | 0.2731 | 0.4908 | 0.6655 |
| | $m = 20$ | 0.0101 | 0.0669 | 0.2402 | 0.4445 | 0.6229 |
| | $m = 25$ | 0.0110 | 0.0629 | 0.2209 | 0.4030 | 0.5783 |

**Table 3:** Power for fixed window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\theta = 2, L = 10000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.0988 | 0.6308 | 0.9609 | 0.9972 | 0.9998 |
| | $m = 7$ | 0.1028 | 0.6785 | 0.9728 | 0.9986 | 0.9999 |
| | $m = 10$ | 0.1018 | 0.7219 | 0.9831 | 0.9993 | 1.0000 |
| | $m = 15$ | 0.1048 | 0.6782 | 0.9738 | 0.9984 | 0.9999 |
| | $m = 20$ | 0.1023 | 0.6427 | 0.9621 | 0.9969 | 0.9996 |
| | $m = 25$ | 0.1048 | 0.6128 | 0.9514 | 0.9950 | 0.9995 |
| $\alpha = 0.05$ | $m = 5$ | 0.0495 | 0.5324 | 0.9366 | 0.9951 | 0.9995 |
| | $m = 7$ | 0.0491 | 0.5836 | 0.9584 | 0.9967 | 0.9998 |
| | $m = 10$ | 0.0517 | 0.6297 | 0.9718 | 0.9983 | 1.0000 |
| | $m = 15$ | 0.0511 | 0.5827 | 0.9586 | 0.9966 | 0.9998 |
| | $m = 20$ | 0.0509 | 0.5435 | 0.9403 | 0.9941 | 0.9994 |
| | $m = 25$ | 0.0487 | 0.5051 | 0.9192 | 0.9907 | 0.9992 |
| $\alpha = 0.01$ | $m = 5$ | 0.0100 | 0.3366 | 0.8563 | 0.9804 | 0.9971 |
| | $m = 7$ | 0.0082 | 0.4008 | 0.9035 | 0.9889 | 0.9991 |
| | $m = 10$ | 0.0110 | 0.4552 | 0.9333 | 0.9948 | 0.9997 |
| | $m = 15$ | 0.0097 | 0.3933 | 0.9008 | 0.9889 | 0.9992 |
| | $m = 20$ | 0.0106 | 0.3387 | 0.8586 | 0.9820 | 0.9988 |
| | $m = 25$ | 0.0083 | 0.2948 | 0.8172 | 0.9703 | 0.9972 |

**Table 4:** Power for fixed window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\alpha = 5, L = 10,000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.1052 | 0.9498 | 0.9999 | 1.0000 | 1.0000 |
| | $m = 7$ | 0.1019 | 0.9699 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 10$ | 0.1069 | 0.9818 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.1051 | 0.9659 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.1035 | 0.9455 | 0.9999 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.1016 | 0.9286 | 0.9997 | 1.0000 | 1.0000 |
| $\alpha = 0.05$ | $m = 5$ | 0.0551 | 0.9173 | 0.9999 | 1.0000 | 1.0000 |
| | $m = 7$ | 0.0527 | 0.9466 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 10$ | 0.0531 | 0.9672 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.0539 | 0.9449 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0534 | 0.9157 | 0.9998 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0521 | 0.8831 | 0.9994 | 1.0000 | 1.0000 |
| $\alpha = 0.01$ | $m = 5$ | 0.0117 | 0.8066 | 0.9994 | 1.0000 | 1.0000 |
| | $m = 7$ | 0.0118 | 0.8709 | 0.9998 | 1.0000 | 1.0000 |
| | $m = 10$ | 0.0126 | 0.9195 | 0.9999 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.0120 | 0.8712 | 0.9997 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0105 | 0.8100 | 0.9988 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0107 | 0.7506 | 0.9972 | 1.0000 | 1.0000 |

**Table 5:** Power for fixed window scan statistic (exponential random variables): $N = 100$, true $m = 10, L = 10,000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.1010 | 0.3878 | 0.7722 | 0.9226 | 0.9767 |
| | $m = 7$ | 0.1020 | 0.4211 | 0.8103 | 0.9406 | 0.9827 |
| | $m = 10$ | 0.1028 | 0.4491 | 0.8352 | 0.9533 | 0.9891 |
| | $m = 15$ | 0.1036 | 0.4194 | 0.8037 | 0.9401 | 0.9831 |
| | $m = 20$ | 0.1022 | 0.4012 | 0.7754 | 0.9255 | 0.9760 |
| | $m = 25$ | 0.1036 | 0.3854 | 0.7494 | 0.9160 | 0.9690 |
| $\alpha = 0.05$ | $m = 5$ | 0.0489 | 0.2893 | 0.6924 | 0.8866 | 0.9613 |
| | $m = 7$ | 0.0504 | 0.3259 | 0.7406 | 0.9118 | 0.9728 |
| | $m = 10$ | 0.0520 | 0.3511 | 0.7775 | 0.9316 | 0.9809 |
| | $m = 15$ | 0.0526 | 0.3204 | 0.7355 | 0.9121 | 0.9727 |
| | $m = 20$ | 0.0518 | 0.2964 | 0.6941 | 0.8927 | 0.9613 |
| | $m = 25$ | 0.0513 | 0.2761 | 0.6633 | 0.8730 | 0.9498 |
| $\alpha = 0.01$ | $m = 5$ | 0.0095 | 0.1459 | 0.5165 | 0.7856 | 0.9088 |
| | $m = 7$ | 0.0106 | 0.1768 | 0.5848 | 0.8320 | 0.9351 |
| | $m = 10$ | 0.0111 | 0.1964 | 0.6346 | 0.8687 | 0.9547 |
| | $m = 15$ | 0.0112 | 0.1648 | 0.5730 | 0.8349 | 0.9383 |
| | $m = 20$ | 0.0118 | 0.1413 | 0.5150 | 0.7916 | 0.9140 |
| | $m = 25$ | 0.0122 | 0.1232 | 0.4635 | 0.7504 | 0.8894 |

## 2.5 Concluding Remarks

In this chapter, conditional fixed window scan statistic has been introduced for detecting a local change in the scale parameter for observation modeled by a gamma distribution with a known shape parameter. Algorithms for implementing the testing procedure based on the scan statistic, as well as evaluating its performance via power calculations, have been developed. Numerical results via Monte Carlo simulation for selected parameters of the models have been presented.

In Section 2.4, Figure 1, for a scanning window of length close to the true length $m$ where a local change in the scale parameter has occurred, with $\theta \geq 1$, the power of test reaches 0.75 or higher when the ratio $\beta_1/\beta_0 = 3$. The simulation results show that when the local change in scale parameter $\beta$ is small, unless the shape parameter $\theta$ is relatively large, the performance of the fixed window scan statistic is not satisfying. When the local change in scale parameter is moderate or large, the algorithm performs well. The power grows higher as the shape parameter $\theta$ becomes larger. Since the scanning window length is determined prior to the calculation and only one window length is considered, the choice of $m$ to be tested is crucial. If the chosen scanning window size is far away from the true size, it could lower the test power substantially. Therefore, if the window size where a potential local change in the scale parameter occurs is unknown, multiple or variable window scan statistics need to be considered. We investigate these type of scan statistics for one dimensional data in Chapter 3.

# Chapter 3

# Multiple and Variable Window Scan Statistics for One Dimensional Gamma Random Variables

## 3.1 Introduction

In Chapter 2, we investigated the performance of a conditional fixed window scan statistic for one dimensional data modeled by a gamma distribution with a known shape parameter and the exponential distribution. In the testing problem we have assumed that length of the sliding window $m$ is known. The simulation study indicated that the test statistic has a higher power when the size of the scanning window is closer to the true window length, where a local change in the scale parameter has occurred. In most applications, $m$ is unknown. Using a fixed window scan statistic in testing the hypotheses specified in (2.2) will result in loss of power.

To address this issue, a possible approach is to scan the data with multiple windows

(Zhao and Glaz, 2016). In addition, to investigate the performance of the scan statistic based on multiple windows, we propose to study the performance of a variable window scan statistic, which will be based on a wide range of possible window sizes. The performance of these statistics is investigated in this chapter and compared to the performance of the fixed window scan statistic, via simulation study. In this chapter we also assume that the shape parameter is known and remains constant for all observations.

This chapter is organized as follows. Section 3.2 presents a conditional multiple window scan statistic for detecting a local change in the scale parameter. Section 3.2.1 discusses the general case for a sequence of gamma random variables, testing the hypotheses stated in (2.2), while Section 3.2.2 is devoted to a conditional multiple window scan statistic for a sequence of exponential random variables, for testing the hypotheses stated in (2.17). To implement these scan statistics for the specified null hypotheses of no local change in the scale parameter, and to evaluate the power under specified alternative hypotheses, simulation algorithms are given in Section 3.2.1. In Section 3.3, the conditional variable window scan statistic for detecting a local change for gamma random variables is developed based on the generalized likelihood ratio method. Methodologies for gamma and exponential random variables are discussed in Sections 3.3.1 and 3.3.2 respectively. In Section 3.4, numerical results from Monte Carlo simulations are presented to evaluate and compare the performance of the conditional fixed window, multiple window and variable window scan statistics. In addition, we present a real world example in Section 3.5, to demonstrate the application of multiple window and variable window

scan statistics, as well as the comparison between the two test statistics. This chapter concludes with a brief discussion in Section 3.6.

## 3.2 Multiple Window Scan Statistics

### 3.2.1 One Dimensional Gamma Random Variables

Let $X_1, X_2, ..., X_N$ be a sequence of independent and identically distributed (i.i.d.) observations from a gamma distribution with shape parameter $\theta$ and scale parameter $\beta$. We are interested in detecting a local change in the scale parameter $\beta$ within a sub-sequence of $m$ consecutive observations in the observed data. While the shape parameter $\theta$ is assumed to be known, the value of $m$ and the starting position are both unknown. The hypotheses to be tested are stated in (2.2). In this chapter, only the detection of an upward shift in $\beta$ is discussed. The methods used in this chapter can be modified to detect either a local downward shift or a two-sided shift.

Although the actual window length is unknown, prior knowledge on the range of the window length can be employed to select a set of $k$ possible window sizes: $2 \leq m_1 < m_2 < ... < m_K \leq N/4$. The corresponding sequence of conditional scan statistics $S^*_{m_1}, ..., S^*_{m_K}$ can be employed simultaneously to detect a local change in the scale parameter $\beta$, where

$$S^*_{m_k} = \max\{Y^*_{j,m_k}; 1 \leq j \leq N - m_k + 1\}, \quad 1 < k < K, \tag{3.1}$$

is defined in (2.9). The moving sums $Y^*_{j,m_k}$ are defined as follows:

$$Y^*_{j,m_k} = \sum_{i=j}^{j+m_k-1} X_i/Y, 1 \leq j \leq N - m_k + 1, \tag{3.2}$$

where $Y$ is the total sum of all observations, defined in Equation (2.4). For $1 \leq k \leq K$, let $t_k$ be the observed value of $S^*_{m_k}$ and $p_k$ be its associated p-value. To test the hypotheses stated in (2.2), we propose to employ the following minimum $p$-value statistic, denoted by $P_{\min}$:

$$P_{\min} = \min\{p_k; 1 \leq k \leq K\}. \tag{3.3}$$

In the context of multiple testing, one can consider the $P_{\min}$ statistic as a nonparametric bootstrap test statistic (Davison and Hinkley, 1997, Sec 4.4.3). The implementation of the $P_{\min}$ test statistic is accomplished via simulation. If for a specified significance level $\alpha$ the null hypothesis specified in (2.2) is rejected, the value of $m_k$ corresponding to the smallest observed $p_k$ value, can be used as an estimate for window size where local change in the scale parameter has occurred.

Algorithm 3 outlined below is used to evaluate the critical value $p^K_\alpha$ for the minimum $p$-value statistic, defined by $P^*_{H_0}(P_{\min} < p^K_\alpha) = \alpha$. The simulated cumulative distribution of $S^*_{m_k}$ can be derived via:

$$P^*(S^*_{m_k} < t_k) = r/R, \quad U^{(r)}_{m_k} \leq t_k < U^{(r+1)}_{m_k}, \tag{3.4}$$

**Algorithm 3:** Multiple Window Scan Statistics: Critical Value for $P_{\min}$

**Result:** Obtain the critical value $p_\alpha^K$ for a given significance level $\alpha$

**for** $r \leftarrow 1$, $R$ **do**
  Draw a sample of $Z_i$'s from $\mathrm{Dir}(\vec{\theta}_N)$ distribution, $i = 1, \ldots, N$;
  **for** $m_k \in \{m_k\}_{k=1}^K$ **do**
    Calculate $N - m_k + 1$ moving sums $Y^*_{j,m_k}$;
    Find the maximum moving sum $S^{(r)*}_{m_k} = \max\{Y^*_{j,m_k}; 1 \le j \le N - m_k + 1\}$, and sort them in ascending order as a vector $U^{(r)}_{m_k}$, thus the simulated cumulative distribution function (3.4) is based on $\{U^{(r)}_{m_k}\}$.
  **end**
**end**
**for** $s \leftarrow 1, S$ **do**
  Simulate a sample of $Z_i$'s from $\mathrm{Dir}(\vec{\theta}_N)$ distribution, $i = 1, \ldots, N$;
  Calculate $S^{(s)}_{m_k}$ for each $m_k$ in the set $\{m_k \mid k = 1, \ldots, K\}$;
  **for** $m_k \in \{m_k\}_{k=1}^K$ **do**
    obtain the $p$-value $p^{(s)}_k$;
  **end**
  Compute the minimum $p$-value $P^{(s)}_{\min}$;
**end**
Sort $\{P^{(s)}_{\min} \mid s = 1, \ldots, S\}$ in ascending order as a vector $Q^{(s)}$, and the critical values $p^K_\alpha$ can be calculated from the simulated $100(1-\alpha)$th percentile of $\{Q^{(s)}\}$.

where $r$ is the number of times out of $R$ trials in the simulation that $S^*_{m_k}$ is less than $t_k$.

The $p$-value $p_k^{(s)}$ is defined as:

$$p_k^{(s)} = P^*(S_{m_k}^{(s)*} > t_k) = r/R, \quad U_{m_k}^{(r)} \leq t_k < U_{m_k}^{(r+1)}, \tag{3.5}$$

where $r$ is the number of times out of $R$ trials in the simulation that $S_{m_k}^{(s)*}$ is larger than $t_k$.

And $P_{\min}^{(s)}$ is defined as:

$$P_{\min}^{(s)} = \min\{p_k^{(s)}; 1 \leq k \leq K\}, \tag{3.6}$$

the simulated cumulative distribution of $P_{\min}$ then can be given by:

$$P^*(P_{\min} < t) = s/S, \quad Q^{(s)} \leq t < Q^{(s+1)}, \tag{3.7}$$

where $Q^{(s)}$ is the sorted vector of $P_{\min}^{(s)}$, and $s$ is the number of times out of $S$ trials in the simulation that $P_{\min}$ is less than $t$.

Based on the simulated cumulative distribution of the minimum $p$-value statistic $P_{\min}$ obtained from Algorithm 3, Algorithm 4, given below, is used to evaluate the power of $P_{\min}$ for specified alternative hypothesis settings. The simulation study presented in Section 3.4 includes results for $\beta_1 \in \{1, 2, 3, 4, 5\}$ and $\beta_0 = 1$.

**Algorithm 4:** Multiple Window Scan Statistics: Power of Test

**Result:** Power of the test for alternative hypothesis set by $\beta_1/\beta_0 = 1, 2, 3, 4, 5$

1 Choose an arbitrary value for the starting position $j$ of the local change, $1 \leq j \leq N - m + 1$, and the total number of simulations $L$ (say $L = 10,000$);
2 **for** $l \leftarrow 1$, $L$ **do**
3  Generate $X_1, ..., X_{j-1} \sim \Gamma(\theta, \beta_0)$, $X_j, ..., X_{j+m-1} \sim \Gamma(\theta, \beta_1)$, and $X_{j+m}, ..., X_N \sim \Gamma(\theta, \beta_0)$;
4  **for** $m_k \in \{m_k\}_{k=1}^{K}$ **do**
5  Calculate $S_{m_k}^{(l)*}$;
6  Compute the associated $p$-value ${p_k}^{(l)}$ for $S_{m_k}^{(l)*}$;
7  **end**
8  Calculate the minimum $p$-value statistic $P_{\min}^{(l)}$;
9  Compare $P_{\min}^{(l)}$ to the simulated cumulative distribution based on $\{Q^{(s)}\}$ obtained from Algorithm 3, and calculate the $p$-value $p^{(l)}$ for this single iteration.
10 **end**
11 Obtain the power $\widehat{\eta}_\alpha$.

The $p$-value in step 9 of Algorithm 4, $p^{(l)}$, is defined as:

$$p^{(l)} = P^*(P_{\min}^{(l)} < t) = s/S, \quad Q^{(s)} \leq t < Q^{(s+1)}, \tag{3.8}$$

where $s$ is the number of times out of $S$ trials in the simulation that $P_{\min}^{(l)}$ is less than $t$. The power of the test statistic $P_{\min}$ , denoted by $\widehat{\eta}_\alpha$, is given by:

$$\widehat{\eta}_\alpha = \frac{\#\{p^{(l)} < \alpha, l = 1, ..., L\}}{L}. \tag{3.9}$$

### 3.2.2 One Dimensional Exponential Random Variables

In the hypothesis testing problem discussed in Section 3.2.1, when $\theta = 1$ the observations $X_1, X_2, ..., X_N$ follow an exponential distribution Exp($\beta$). We are interested in testing the hypotheses stated in (2.17). As in Section 3.2.1, we are only investigating a local upward shift.

When the window length $m$, where a local change in the scale parameter occurs is unknown, we employ a conditional multiple window scan statistic based the on the $P_{\min}$ statistic as we discussed in the section 3.2.1. Based on prior experience, a set of possible window lengths $\{m_k\}_{k=1}^{K}$ for the occurrence of a local change in the scale parameter is selected. The maximum moving sum $S^*_{m_k}$ as well as the $p$-value are subsequently calculated for each fixed window of size $m_k$. Then, the minimum $p$-value statistic $P_{\min}$, defined in (3.3), is employed to test the hypotheses. Algorithms 3 and 4 can be adopted to calculate the critical value of $P_{\min}$ and the power of test, by setting $\theta = 1$.

## 3.3 Variable Window Scan Statistics

### 3.3.1 One Dimensional Gamma Random Variables

In the previous section, a multiple window scan statistic, based on the $P_{\min}$ statistic has been discussed. We now present a different approach for testing hypotheses stated in (2.2), based on the generalized likelihood ratio method. For Bernoulli and Poisson

models, the generalized likelihood ratio approach for unconditional variable window scan statistics was introduced in Kulldorff (1997) and Nagawalla (1996).

Let $X_1, X_2, ..., X_N$ be a sequence of i.i.d. observations from a gamma distribution $\Gamma(\theta, \beta)$. Our goal is to detect a local upward shift in the scale parameter $\beta$ within a sub-sequence of $m$ consecutive observations in the observed data, while neither the value of $m$ nor the starting position $j$ of the local shift are known. We are interested in testing:

$$H_0 : X_i \sim \Gamma(\theta, \beta_0), \forall i = 1, 2, ..., N; X_i\text{'s are independent;} \quad \text{vs.}$$

$$H_a : \exists j, 1 \leq j \leq N - m + 1, \text{s.t. } X_i \sim \Gamma(\theta, \beta_1), \beta_1 > \beta_0, \forall i = j, ..., j + m - 1;$$

$$\text{and } X_i \sim \Gamma(\theta, \beta_0), \forall i = 1, ...j - 1, j + m, ..., N; X_i\text{'s are independent;}$$

where $\beta_0$ and $\beta_1$ are unknown. Assume that the true window size $m$ is between $m_0$ and $\widetilde{m}_0$, where $3 \leq m_0 \leq m \leq \widetilde{m}_0 \leq N/4$. Since the scale parameter $\beta_0$ under null is unknown, we derive below a test statistics via the generalized likelihood ratio principle, by conditioning on both the sum of all the observation is the data, $Y = \sum_{i=1}^{N} X_i$, and the sum of the partial data, $Y_{j,m} = \sum_{i=j}^{j+m-1} X_i$, where $3 \leq m \leq N/4$, corresponding to a specified alternative hypothesis. We refer to this test statistic as a *conditional variable window scan statistic*.

For the problem at hand, the conditional general likelihood ratio test (cGLRT) is

given by:

$$\begin{aligned}
\Lambda &= \frac{\sup_{\Theta_1} f(x_1,...,x_N|y,y_{j,m})}{\sup_{\Theta_0} f(x_1,...,x_N|y,y_{j,m})} \\
&= \frac{\sup_{\Theta_1}\left\{\frac{\Gamma(m\theta)\Gamma[(N-m)\theta]}{[\Gamma(\theta)^N]}\cdot\frac{[\prod_{i=j+1}^{j+m-1}x_i\cdot(y_{j,m}-\sum_{j+1}^{j+m-1}x_i)\cdot\prod_{i\in I}x_i\cdot(y-y_{j,m}-\sum_{i\in I}x_i)]^{\theta-1}}{y_{j,m}^{m\theta-1}\cdot(y-y_{j,m})^{(N-m)\theta-1}}\right\}}{\sup_{\Theta_0}\left\{\frac{\Gamma(N\theta)}{[\Gamma(\theta)]^N}\cdot\frac{[\prod_{i=1}^{N-1}x_i\cdot(y-\sum_{i=1}^{N-1}x_i)]^{\theta-1}}{y^{N\theta-1}}\right\}} \\
&= \sup_{\Theta_1}\left\{\frac{\Gamma(m\theta)\Gamma[(N-m)\theta]}{\Gamma(N\theta)}\cdot\frac{y^{N\theta-1}}{y_{j,m}^{m\theta-1}(y-y_{j,m})^{(N-m)\theta-1}}\right. \\
&\qquad\left.\cdot\left[\frac{(y_{j,m}-\sum_{j+1}^{j+m-1}x_i)(y-y_{j,m}-\sum_{i\in I}x_i)}{x_{j+m}(y-\sum_{i=1}^{N-1}x_i)}\right]^{\theta-1}\right\},
\end{aligned}$$

where $f(x_1,...,x_N|y,y_{j,m})$ is the joint distribution of $X_1,\ldots,X_N$ conditional on the total sum $Y$ and the partial sum $Y_{j,m}$, $\Theta_0$ and $\Theta_1$ denote the respective parameter spaces under $H_0$ and $H_1$, and $I=\{1,\ldots,j,j+m+1,\ldots,N-1\}$ is the index set. Since the last term in the equation above equals 1, it can be simplified as:

$$\begin{aligned}
\Lambda &= \Lambda(j,m \mid y,y_{j,m}) \\
&= \sup_{\Theta_1}\left\{\frac{\Gamma(m\theta)\Gamma[(N-m)\theta]}{\Gamma(N\theta)}\cdot\frac{y^{N\theta-1}}{y_{j,m}^{m\theta-1}(y-y_{j,m})^{(N-m)\theta-1}}\right\} \\
&= \sup_{j,m}\left\{\frac{\Gamma(m\theta)\Gamma[(N-m)\theta]}{\Gamma(N\theta)}\cdot\frac{y^{N\theta-1}}{y_{j,m}^{m\theta-1}(y-y_{j,m})^{(N-m)\theta-1}}\right\},
\end{aligned}$$

which can further be simplified as:

$$\Lambda = \sup_{j,m}\left\{\frac{\Gamma(m\theta)\Gamma[(N-m)\theta]}{\Gamma(N\theta)}\cdot\frac{y}{(y_{j,m}/y)^{m\theta-1}(1-y_{j,m}/y)^{(N-m)\theta-1}}\right\}. \tag{3.10}$$

To implement the test statistic $\Lambda$, we define the function

$$g(u) = u^{1-m\theta}(1-u)^{1-(N-m)\theta}, \quad 0 < u < 1, \tag{3.11}$$

where $u = y_{j,m}/y$. Hence, we get that:

$$\Lambda = \sup_{j,m} \left\{ \frac{\Gamma(m\theta)\Gamma[(N-m)\theta]y}{\Gamma(N\theta)} \cdot g(u) \right\}. \tag{3.12}$$

For the conditional generalized likelihood ratio test, the null hypothesis is rejected for large values of $\Lambda$. To obtain the observed value of $\Lambda$, we need to evaluate $g(u)$. Since the first term of $\Lambda$ is a constant, to determine the maximum we only need to find $\max\{g(u)\}$.

For a fixed value of $m$, since $g(u)$ is a convex function of of $u$, the maximum value of the function should be obtained at either the smallest or the largest value of $u$, i.e. the smallest or largest value of value of $y_{j,m}/y$. Furthermore, since in our alternative hypothesis $\beta_1 > \beta_0$, which indicates the expectation of the observations within the local change window being larger than the remaining $X_i$'s. Under the circumstances, a small ratio of $y_{j,m}/y$ would not lead to a conclusion in favor of $H_a$. We only need to find $\max\{y_{j,m}/y\}$, i.e. the maximum moving sum $S_m^*$ in the sample and calculate the corresponding likelihood ratio, denoted by $L^*(m)$:

$$L^*(m) = \frac{\Gamma(m\theta)\Gamma[(N-m)\theta]}{\Gamma(N\theta)} \cdot \frac{y}{(S_m^*)^{m\theta-1}(1-S_m^*)^{(N-m)\theta-1}}. \tag{3.13}$$

The next step is to compute $L^*(m)$ for all values of $m$ between $m_0$ and $\widetilde{m}_0$, which in the most complete case would be all the integers between 3 and $N/4$. Then, we can obtain the observed value of the variable window scan statistic $\Lambda^*$ as:

$$\Lambda^* = \max\{L^*(m) \mid m_0 \leq m \leq \widetilde{m}_0\}. \tag{3.14}$$

We can denote the corresponding $m$ as $m^*$, which is the mostly likely size of the window where the local change occurs. In addition, we can also record the location $j$ of the maximum moving sum $S^*_{m^*}$ as $j^*(m^*)$, that is the most likely starting position of the local change.

In Algorithm 5, we outline the steps for implementing the variable window scan statistic. A Monte Carlo simulation can be used to find the simulated cumulative distribution of this test statistic, along with the critical value $p^{\Lambda}_{\alpha}$, which we present in Algorithm 6. $\Phi^{(r)}$, which we can obtain from Algorithm 6 can be used to construct the simulated cumulative distribution of the variable window scan statistic $\Lambda^*$:

$$P^*(\Lambda^* < t) = r/R, \quad \Phi^{(r)} \leq t < \Phi^{(r+1)}, \tag{3.15}$$

where $r$ is the number of times out of $R$ trials in the simulation that $\Lambda^*$ is less than $t$.

The power of the variable window scan statistic can also be calculated in a similar way to the fixed window scan statistic, via Monte Carlo simulation, as demonstrated in

**Algorithm 5:** Variable Window Scan Statistic: Maximum Likelihood Ratio

**Result:** The maximum likelihood ratio for a given range of window lengths

**for** $m \leftarrow m_0, \widetilde{m}_0$ **do**

  Compute moving sums $Y^*_{j,m}$ for each $j$ that $1 \leq j \leq N - m + 1$;

  Find the maximum moving sum $S^*_m$, and record the corresponding $j$ recorded as $j^*(m)$;

  Calculate $L^*(m)$ by Equation (3.13)

**end**

Find $\Lambda^*$ as defined in (3.14), as well as the corresponding $m^*$ and $j^*(m^*)$.

**Algorithm 6:** Variable Window Scan Statistic: Critical Value of $\Lambda^*$

**Result:** The critical value for the generalized likelihood ratio

**for** $r \leftarrow 1, R$ **do**

  Simulate a sample of $Z_i$'s that follow $\text{Dir}(\vec{\theta}_N)$ distribution, $i = 1, ..., N$;

  Calculate the variable window scan statistic $\Lambda^{*(r)}$ as defined in 3.14, and sort them in increasing order as a vector $\Phi^{(r)}$. The simulated cumulative distribution function (3.15) is based on $\Phi^{(r)}$;

**end**

The critical values $p^{\Lambda}_{\alpha}$ can be calculated from the simulated $100(1-\alpha)$th percentile of $\Phi^{(r)}$.

Algorithm 2. Numerical results for the power of the variable window scan statistic, for selected values of the parameters, are presented in Section 3.4.

### 3.3.2 One Dimensional Exponential Random Variables

We now consider the special case when $X_1, X_2, ..., X_N$ is a sequence of i.i.d. observations from the exponential distribution Exp($\beta_0$). We are interested in detecting a upward shift in $\beta_0$, i.e. an occurrence of a sub-sequence of observations following Exp($\beta_1$), $\beta_1 > \beta_0$, while the window length $m$ of this local change is unknown. The hypotheses to be tested is (2.17).

Suppose that based on prior experience, one can assume that the value of $m$ is: $3 \leq m_0 \leq m \leq \widetilde{m}_0 \leq N/4$. We can calculate the variable window scan statistic following the methodology discussed in Section 3.3.1, setting $\theta = 1$, and calculate the $p$-value for the hypothesis test. If no accurate guess of $m$ can be made, one can simply use all possible values from $m = 3$ to $N/4$. If the observed $p$-value $< \alpha$, then the null hypothesis is rejected. In this case a local change in the scale parameter has occurred and one can also estimate the most likely window length and the starting position of the local change. Numerical results for the exponential model are presented in Section 3.4.

## 3.4 Numerical Results: Simulation

In Sections 3.2 and 3.3, we introduced conditional multiple window and conditional variable window scan statistics, respectively. In this section we present numerical results for both scan statistics based on simulation algorithms developed in previous sections. We compare the performance of these scan statistics with the conditional fixed window scan statistic discussed in Chapter 2, by comparing the power for selected parameters of the model. In Figure 2, we present the power for conditional multiple window and variable window scan statistics, as well as that of the conditional fixed window scan statistic. These comparisons are all based on samples generated with a local change of in a sliding window of length $m = 10$, within a sequence of $N = 100$ observations, with different ratios of $\beta_1/\beta_0 \in \{1, 2, 3, 4, 5\}$. The power of these scan statistics is calculated from $L = 10,000$ Monte Carlo simulations, for significance levels $\alpha = 0.10$, 0.05 and 0.01. The corresponding power values are recorded in Tables 7 to 10. Tables 7, 8 and 9 present results for gamma random variables, with $\theta = 0.5$, 2, and 5 respectively, and Table 10 is for exponential random variables. Note that in these comparisons, the accuracy of detecting the location of the local change is not taken into consideration.

Abbreviations are summarized in Table 6 for reference. For different approaches mentioned, "FW(T)" stands for fixed window scan statistic with the correct window size of local change ($m = 10$), "FW(F)" is for fixed window scan statistic with a wrong window size ($m = 5$), "MW(I)" indicates multiple window scan statistic with the window

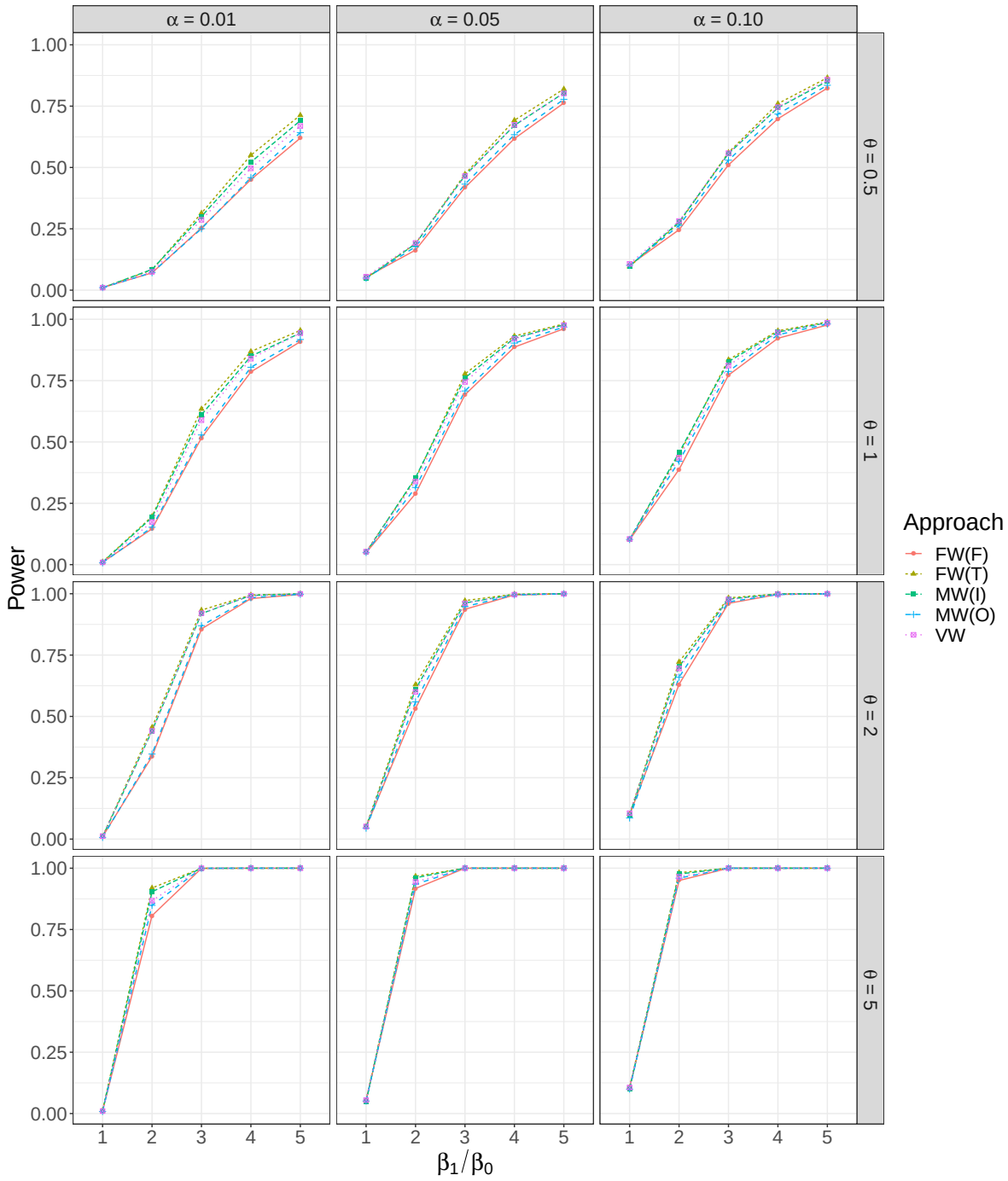


**Figure 2:** Power comparison for different scan statistics (gamma and exponential random variables): $N = 100$, true $m = 10$, $\theta = 0.5, 1, 2$ and 5

sizes tested including $m = 10$ ($m_k = 5, 10$ and 20), "MW(O)" means multiple window scan statistic with the true window size falling outside of the tested window lengths ($m_k = 15$, 20, and 25 ), and "VW" for variable window scan statistic with a testing range of $m$ from 3 to 25.

In Figure 2, the four rows of plots demonstrate power results for gamma random variables with shape parameter $\theta = 0.5$, 1, 2 and 5. Note that gamma random variables with $\theta = 1$ are exponential random variables. We present these results in the current order only for the convenience of showing the trend for increasing values of $\theta$. The three columns of plots correspond to significance level $\alpha = 0.01$, 0.05 and 0.10 from right to left.

When the local shift of $\beta_1/\beta_0$ in the alternative hypothesis is small, none of the proposed scan statistics show high power in detecting the local change. But as the local shift of $\beta_1/\beta_0$ becomes larger, all of the three approaches perform well except for the fixed window scan statistics using an incorrect window size (far away from true size). The power of these approaches can be ranked from in descending order: FW(T) $>$ MW(I) $\approx$ VW $>$ MW(O) $>$ FW(F). The rankings are also presented in Table 6.

**Table 6:** Approach Abbreviations

| Approach | Scan Statistic | True $m$ | $m$ tested | Power Rank |
|---|---|---|---|---|
| FW(T) | Fixed window scan statistic | 10 | 10 | 1 |
| FW(F) | Fixed window scan statistic | 10 | 5 | 5 |
| MW(I) | Multiple window scan statistic | 10 | 5, 10, 20 | 2 (tie) |
| MW(O) | Multiple window scan statistic | 10 | 15, 20, 25 | 4 |
| VW | Variable window scan statistic | 10 | 3 - 25 | 2 (tie) |

The fixed window scan statistic using the exact true window size shows the highest power, as well as the best computational efficiency, because it only scans one window size. Multiple and variable window scan statistics that cover the true window size show slightly lower power but still both perform well, and the difference between these two is not substantial, usually around 0.02 or smaller. The slight decrease in power is expected, since both of them take multiple window sizes into consideration and the probability of making an incorrect conclusion would increase. Multiple window scan statistic with the true window size falling outside of the tested set show further lower power than these three, and the powers of fixed window scan statistic using an incorrect window size are the lowest.

For all the tests, given the same $\beta_1/\beta_0$ and significance level $\alpha$, simulations with larger shape parameter $\theta$ yield higher power, possibly due to $E(X_i) = \theta\beta$, indicating that larger $\theta$ would lead to a larger local shift for observations inside the window.

In Figure 3, we present the power comparison for multiple window scan statistic. The true window lengths tested are $m = 5$, 7, 10, 15, 20 and 25 shown in different colors, and for all of them we employed scanning windows of size $m_1 = 5$, $m_2 = 10$ and $m_3 = 20$. There is a consistent trend across almost all $\beta_1/\beta_0$ ratio, $\theta$ and significance level settings that the power increases as the true window size $m$ becomes larger. Corresponding numerical results can be found in Tables 11 to 14: Tables 11, 12 and 13 are results for gamma random variables, with $\theta = 0.5, 2$, and 5 respectively, and Table 14 is for exponential random variables.

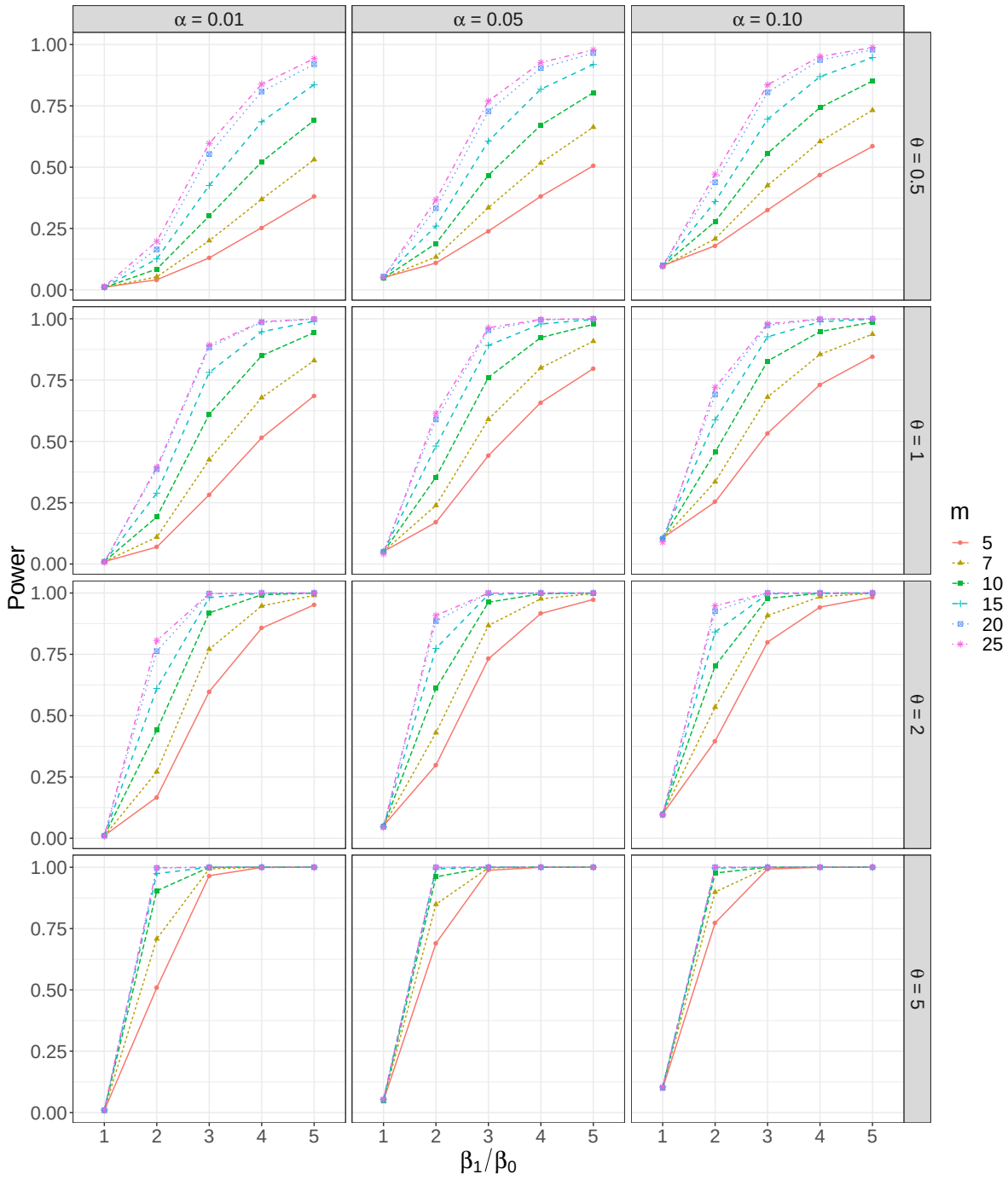


**Figure 3:** Power comparison for multiple window scan statistic (gamma and exponential random variables): $N = 100$, true window lengths $m = 5$, 7, 10, 15, 20 and 25, window lengths tested $m = 5$, 10 and 20, $\theta = 0.5, 1, 2$ and 5

These results are expected for true window lengths of $m = 5$, 10 and 20, since they are all taken into consideration for testing, and as the length of the sliding window where a local change has occurred becomes wider, it is easier to detect. On the other hand, it may be unexpected that the results for true window sizes $m = 7$, 15 and 25 can still align "naturally" in the trend with the other three. A better explanation is provided in another set of comparisons, and the results are presented in Figure 4, where the powers are included for fixed, multiple and variable window scan statistics for true window lengths $m = 7$, 15 and 25. For each of the $m$ value, fixed window scan statistic (FW) is tested with the correct window size, providing the highest power for each comparison, multiple window scan statistic (MW) includes $m_1 = 5$, $m_2 = 10$ and $m_3 = 20$, and variable window scan statistic (VW) tests the range of $m$ from 3 to 25. For $m = 7$, the power difference among the three statistics is very small, mostly under 0.03, and there is no notable difference between MW and VW. This is possibly because $m_1 < 7 < m_2$, and the gap in between is relatively small. For $m = 15$, there is still no distinguishable difference between MW and VW, but slightly larger difference comparing to FW. For $m = 25$, larger difference is shown between FW and MW/VW. It turns out that for multiple window scan statistic, when the difference between the true window size and the tested $\{m_k\}$ set is relatively small, the power decrease from FW with correct $m$ or VW is not substantial, which is consistent from what we can see for MW(O) in Figure 2. However, the power decrease from FW with correct window size to MW/VW becomes larger when $m$ increases. In this comparison, we can also see that with the true window

size, the test power increases as window size becomes larger for the fixed window scan statistic, which is consistent with MW/VW.

Next in Figure 5 we show the power comparison for variable window scan statistic. The true window lengths tested are $m = 5$, 7, 10, 15, 20 and 25 shown in different colors, and for all of them we employed scanning windows of size $m = 3$ to 25. We can validate that, with all true window sizes included in the testing range, the power increases as the true $m$ increases. Corresponding numerical results can be found in Tables 15 to 18, where Tables 15, 16 and 17 are results for gamma random variables, with $\theta = 0.5$, 2, and 5 respectively, and Table 18 is for exponential random variables.

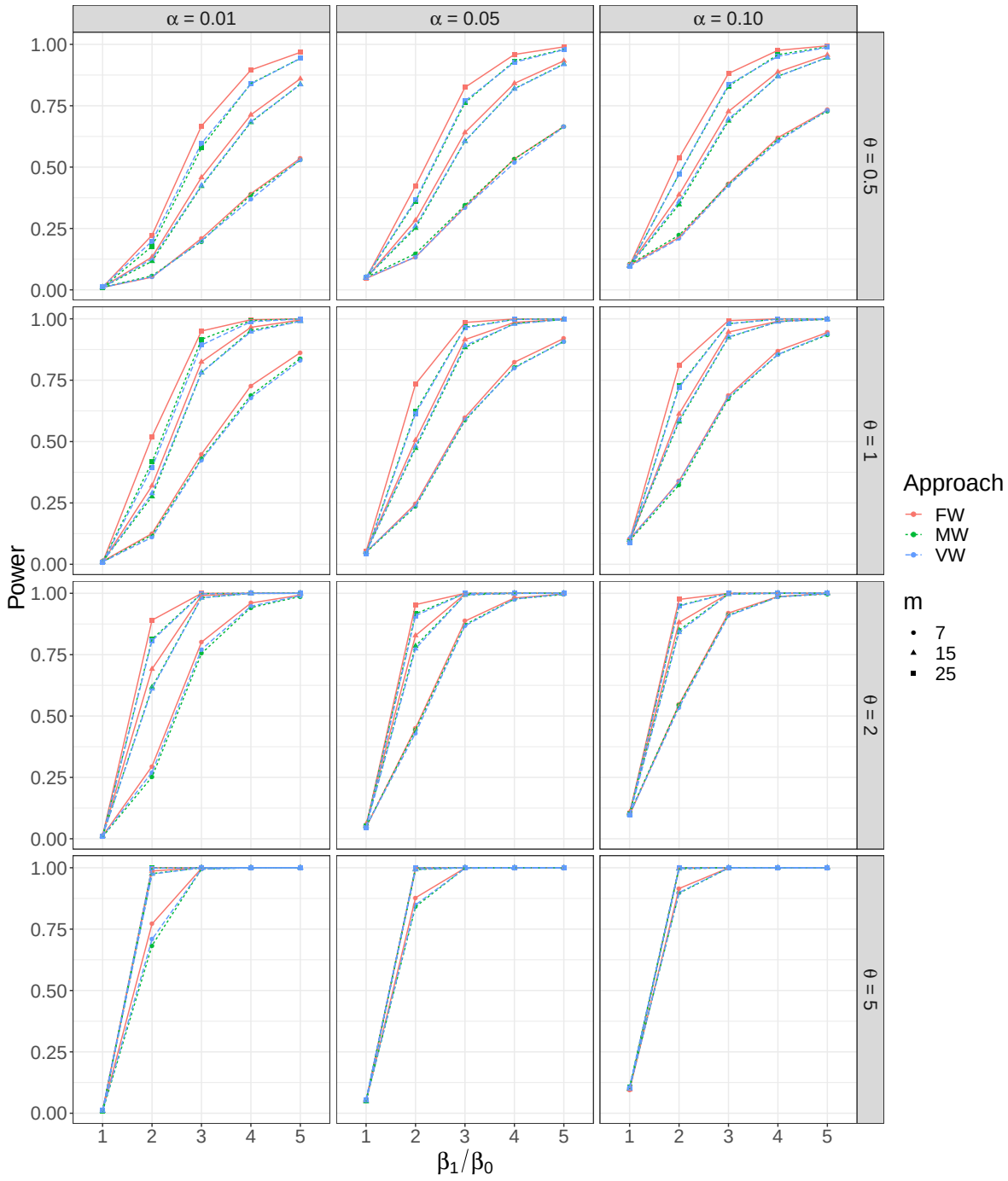


**Figure 4:** Power comparison for different scan statistic (gamma and exponential random variables): $N = 100$, true $m = 7$, 15 and 25, window lengths tested in FW are the true $m$ values, window lengths tested in MW are $m = 5$, 10 and 20, window lengths tested in VM are $m = 3$ to 25, $\theta = 0.5, 1, 2$ and 5

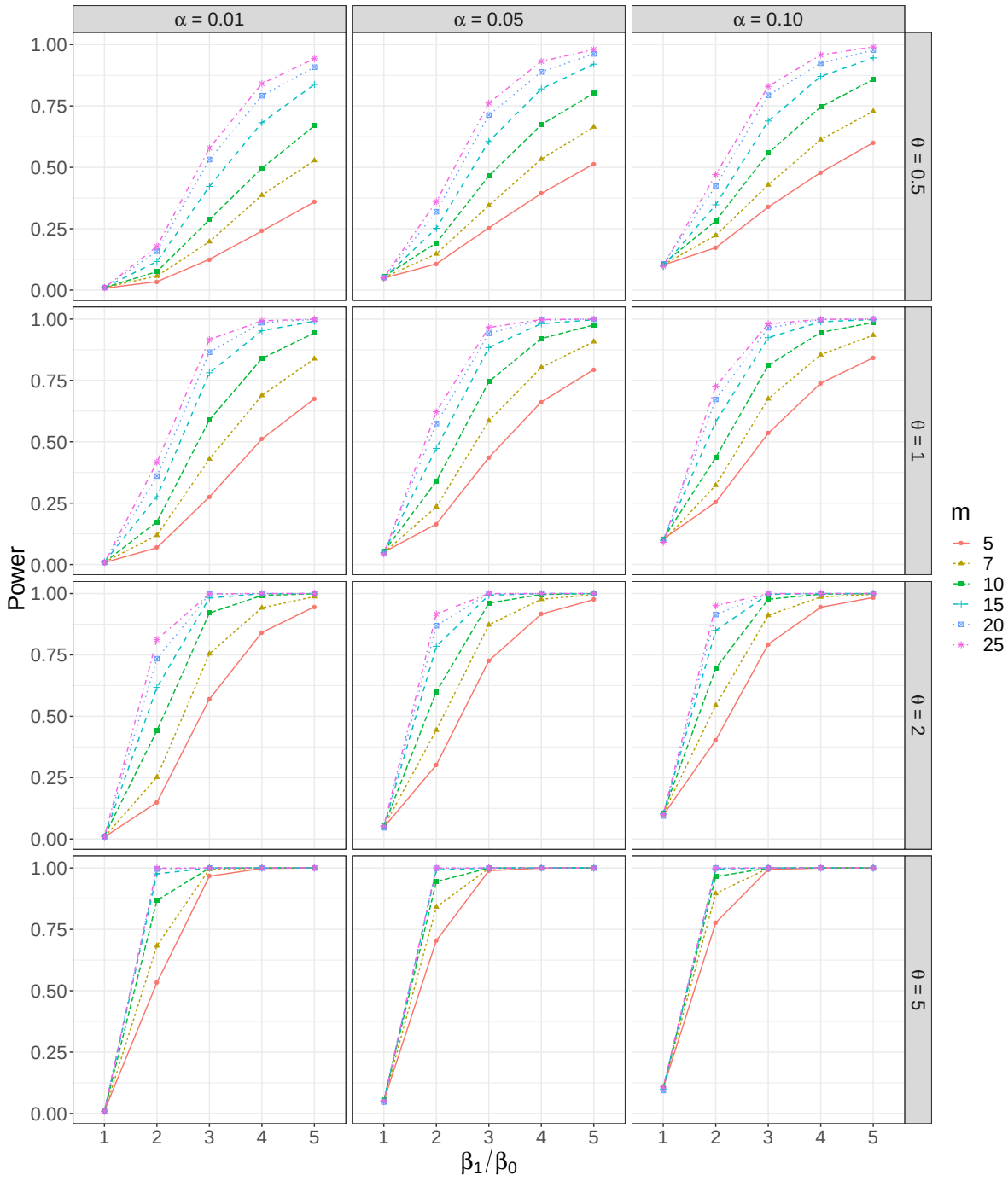


**Figure 5:** Power comparison for variable window scan statistic (gamma and exponential random variables): $N = 100$, true $m = 10$, window lengths tested $m = 3$ to 25, $\theta = 0.5, 1, 2$ and 5

**Table 7:** Power comparison for gamma random variables: $N = 100$, true $m = 10$, $\theta = 0.5, L = 10000$

| | Approach | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | FW(T) | 0.0963 | 0.2745 | 0.5613 | 0.7603 | 0.8658 |
| | FW(F) | 0.1019 | 0.2462 | 0.5104 | 0.6986 | 0.8235 |
| | MW(I) | 0.0975 | 0.2781 | 0.5566 | 0.7431 | 0.8517 |
| | MW(O) | 0.1011 | 0.2637 | 0.5298 | 0.7176 | 0.8351 |
| | VW | 0.1071 | 0.2818 | 0.5580 | 0.7459 | 0.8579 |
| $\alpha = 0.05$ | FW(T) | 0.0492 | 0.1896 | 0.4728 | 0.6924 | 0.8198 |
| | FW(F) | 0.0534 | 0.1634 | 0.4182 | 0.6176 | 0.7629 |
| | MW(I) | 0.0475 | 0.1890 | 0.4672 | 0.6714 | 0.8038 |
| | MW(O) | 0.0490 | 0.1772 | 0.4321 | 0.6336 | 0.7779 |
| | VW | 0.0551 | 0.1923 | 0.4649 | 0.6736 | 0.8018 |
| $\alpha = 0.01$ | FW(T) | 0.0104 | 0.0826 | 0.3140 | 0.5503 | 0.7131 |
| | FW(F) | 0.0106 | 0.0706 | 0.2538 | 0.4508 | 0.6198 |
| | MW(I) | 0.0095 | 0.0847 | 0.3007 | 0.5216 | 0.6910 |
| | MW(O) | 0.0105 | 0.0713 | 0.2506 | 0.4586 | 0.6421 |
| | VW | 0.0103 | 0.0748 | 0.2867 | 0.4966 | 0.6692 |

**Table 8:** Power comparison for gamma random variables: $N = 100$, true $m = 10$, $\theta = 2, L = 10000$

| | Approach | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | FW(T) | 0.1018 | 0.7219 | 0.9831 | 0.9993 | 1.0000 |
| | FW(F) | 0.0988 | 0.6308 | 0.9609 | 0.9972 | 0.9998 |
| | MW(I) | 0.0991 | 0.7032 | 0.9781 | 0.9988 | 0.9999 |
| | MW(O) | 0.0861 | 0.6597 | 0.9669 | 0.9970 | 0.9995 |
| | VW | 0.1054 | 0.6954 | 0.9765 | 0.9976 | 0.9998 |
| $\alpha = 0.05$ | FW(T) | 0.0517 | 0.6297 | 0.9718 | 0.9983 | 1.0000 |
| | FW(F) | .0495 | 0.5324 | 0.9366 | 0.9951 | 0.9995 |
| | MW(I) | 0.0492 | 0.6110 | 0.9624 | 0.9969 | 0.9998 |
| | MW(O) | 0.0445 | 0.5592 | 0.9476 | 0.9952 | 0.9995 |
| | VW | 0.0514 | 0.6001 | 0.9609 | 0.9963 | 0.9998 |
| $\alpha = 0.01$ | FW(T) | 0.0110 | 0.4552 | 0.9333 | 0.9948 | 0.9997 |
| | FW(F) | 0.0100 | 0.3366 | 0.8563 | 0.9804 | 0.9971 |
| | MW(I) | 0.0101 | 0.4408 | 0.9190 | 0.9930 | 0.9995 |
| | MW(O) | 0.0069 | 0.3465 | 0.8695 | 0.9837 | 0.9979 |
| | VW | 0.0121 | 0.4412 | 0.9206 | 0.9921 | 0.9993 |

**Table 9:** Power comparison for gamma random variables: $N = 100$, true $m = 10$, $\theta = 5, L = 10000$

| | Approach | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | FW(T) | 0.1069 | 0.9818 | 1.0000 | 1.0000 | 1.0000 |
| | FW(F) | 0.1052 | 0.9498 | 0.9999 | 1.0000 | 1.0000 |
| | MW(I) | 0.1019 | 0.9763 | 1.0000 | 1.0000 | 1.0000 |
| | MW(O) | 0.0995 | 0.9594 | 1.0000 | 1.0000 | 1.0000 |
| | VW | 0.1061 | 0.9650 | 1.0000 | 1.0000 | 1.0000 |
| $\alpha = 0.05$ | FW(T) | 0.0531 | 0.9672 | 1.0000 | 1.0000 | 1.0000 |
| | FW(F) | 0.0551 | 0.9173 | 0.9999 | 1.0000 | 1.0000 |
| | MW(I) | 0.0483 | 0.9609 | 1.0000 | 1.0000 | 1.0000 |
| | MW(O) | 0.0514 | 0.9345 | 1.0000 | 1.0000 | 1.0000 |
| | VW | 0.0547 | 0.9443 | 0.9999 | 1.0000 | 1.0000 |
| $\alpha = 0.01$ | FW(T) | 0.0126 | 0.9195 | 0.9999 | 1.0000 | 1.0000 |
| | FW(F) | 0.0117 | 0.8066 | 0.9994 | 1.0000 | 1.0000 |
| | MW(I) | 0.0096 | 0.9040 | 0.9997 | 1.0000 | 1.0000 |
| | MW(O) | 0.0085 | 0.8486 | 0.9996 | 1.0000 | 1.0000 |
| | VW | 0.0102 | 0.8677 | 0.9999 | 1.0000 | 1.0000 |

**Table 10:** Power comparison for exponential random variables: $N = 100$, true $m = 10$, $L = 10000$

| | Approach | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | FW(T) | 0.1028 | 0.4491 | 0.8352 | 0.9533 | 0.9891 |
| | FW(F) | 0.1010 | 0.3878 | 0.7722 | 0.9226 | 0.9767 |
| | MW(I) | 0.1028 | 0.4564 | 0.8274 | 0.9478 | 0.9862 |
| | MW(O) | 0.1025 | 0.4214 | 0.7875 | 0.9369 | 0.9807 |
| | VW | 0.1046 | 0.4362 | 0.8115 | 0.9457 | 0.9859 |
| $\alpha = 0.05$ | FW(T) | 0.0520 | 0.3511 | 0.7775 | 0.9316 | 0.9809 |
| | FW(F) | 0.0489 | 0.2893 | 0.6924 | 0.8866 | 0.9613 |
| | MW(I) | 0.0494 | 0.3545 | 0.7611 | 0.9234 | 0.9772 |
| | MW(O) | 0.0499 | 0.3145 | 0.7077 | 0.9028 | 0.9684 |
| | VW | 0.0528 | 0.3391 | 0.7446 | 0.9204 | 0.9762 |
| $\alpha = 0.01$ | FW(T) | 0.0111 | 0.1964 | 0.6346 | 0.8687 | 0.9547 |
| | FW(F) | 0.0095 | 0.1459 | 0.5165 | 0.7856 | 0.9088 |
| | MW(I) | 0.0102 | 0.1920 | 0.6107 | 0.8495 | 0.9439 |
| | MW(O) | 0.0074 | 0.1524 | 0.5285 | 0.8044 | 0.9174 |
| | VW | 0.0092 | 0.174 | 0.5888 | 0.8394 | 0.9434 |

**Table 11:** Power for multiple window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\theta = 0.5, L = 10000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.0987 | 0.1795 | 0.3256 | 0.4691 | 0.5844 |
| | $m = 7$ | 0.0957 | 0.2078 | 0.4254 | 0.6047 | 0.7322 |
| | $m = 10$ | 0.0975 | 0.2781 | 0.5566 | 0.7431 | 0.8517 |
| | $m = 15$ | 0.0970 | 0.3596 | 0.6955 | 0.8692 | 0.9468 |
| | $m = 20$ | 0.1001 | 0.4381 | 0.8051 | 0.9369 | 0.9788 |
| | $m = 25$ | 0.0955 | 0.4715 | 0.8358 | 0.9513 | 0.9880 |
| $\alpha = 0.05$ | $m = 5$ | 0.0498 | 0.1092 | 0.2396 | 0.3810 | 0.5057 |
| | $m = 7$ | 0.0478 | 0.1341 | 0.3348 | 0.5172 | 0.6632 |
| | $m = 10$ | 0.0475 | 0.1890 | 0.4672 | 0.6714 | 0.8038 |
| | $m = 15$ | 0.0495 | 0.2598 | 0.6064 | 0.8182 | 0.9186 |
| | $m = 20$ | 0.0541 | 0.3324 | 0.7278 | 0.9040 | 0.9651 |
| | $m = 25$ | 0.0511 | 0.3671 | 0.7693 | 0.9269 | 0.9780 |
| $\alpha = 0.01$ | $m = 5$ | 0.0110 | 0.0410 | 0.1302 | 0.2534 | 0.3799 |
| | $m = 7$ | 0.0098 | 0.0528 | 0.2005 | 0.3689 | 0.5309 |
| | $m = 10$ | 0.0095 | 0.0847 | 0.3007 | 0.5216 | 0.6910 |
| | $m = 15$ | 0.0100 | 0.1260 | 0.4254 | 0.6851 | 0.8363 |
| | $m = 20$ | 0.0116 | 0.1647 | 0.5532 | 0.8084 | 0.9196 |
| | $m = 25$ | 0.0131 | 0.1977 | 0.5967 | 0.8384 | 0.9428 |

**Table 12:** Power for multiple window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\theta = 2, L = 10000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.0991 | 0.3965 | 0.7999 | 0.9425 | 0.9829 |
| | $m = 7$ | 0.0992 | 0.5345 | 0.9081 | 0.9850 | 0.9986 |
| | $m = 10$ | 0.0991 | 0.7032 | 0.9781 | 0.9988 | 0.9999 |
| | $m = 15$ | 0.0971 | 0.8411 | 0.9966 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0956 | 0.9259 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0959 | 0.9479 | 0.9999 | 1.0000 | 1.0000 |
| $\alpha = 0.05$ | $m = 5$ | 0.0511 | 0.2985 | 0.7311 | 0.9163 | 0.9737 |
| | $m = 7$ | 0.0494 | 0.4306 | 0.8675 | 0.9764 | 0.9973 |
| | $m = 10$ | 0.0492 | 0.6110 | 0.9624 | 0.9969 | 0.9998 |
| | $m = 15$ | 0.0485 | 0.7742 | 0.9937 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0469 | 0.8852 | 0.9996 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0442 | 0.9085 | 0.9996 | 1.0000 | 1.0000 |
| $\alpha = 0.01$ | $m = 5$ | 0.0111 | 0.1673 | 0.5980 | 0.8567 | 0.9516 |
| | $m = 7$ | 0.0096 | 0.2710 | 0.7710 | 0.9472 | 0.9904 |
| | $m = 10$ | 0.0101 | 0.4408 | 0.9190 | 0.9930 | 0.9995 |
| | $m = 15$ | 0.0108 | 0.6110 | 0.9816 | 0.9995 | 0.9999 |
| | $m = 20$ | 0.0107 | 0.7627 | 0.9975 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0097 | 0.8064 | 0.9974 | 1.0000 | 1.0000 |

**Table 13:** Power for multiple window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\theta = 5, L = 10000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.0988 | 0.7733 | 0.9928 | 0.9998 | 1.0000 |
| | $m = 7$ | 0.1020 | 0.8989 | 0.9987 | 1.0000 | 1.0000 |
| | $m = 10$ | 0.1019 | 0.9763 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.1047 | 0.9965 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0999 | 0.9999 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.1040 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| $\alpha = 0.05$ | $m = 5$ | 0.0481 | 0.6901 | 0.9874 | 0.9998 | 1.0000 |
| | $m = 7$ | 0.0513 | 0.8489 | 0.9975 | 1.0000 | 1.0000 |
| | $m = 10$ | 0.0483 | 0.9609 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.0544 | 0.9932 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0519 | 0.9996 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0536 | 0.9999 | 1.0000 | 1.0000 | 1.0000 |
| $\alpha = 0.01$ | $m = 5$ | 0.0087 | 0.5085 | 0.9656 | 0.9991 | 1.0000 |
| | $m = 7$ | 0.0099 | 0.7082 | 0.9937 | 1.0000 | 1.0000 |
| | $m = 10$ | 0.0096 | 0.9040 | 0.9997 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.0090 | 0.9741 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0087 | 0.9969 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0114 | 0.9980 | 1.0000 | 1.0000 | 1.0000 |

**Table 14:** Power for multiple window scan statistic (exponential random variables): $N = 100$, true $m = 10$, L = 10000

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.1064 | 0.2528 | 0.5334 | 0.7315 | 0.8467 |
| | $m = 7$ | 0.1023 | 0.3354 | 0.6810 | 0.8550 | 0.9373 |
| | $m = 10$ | 0.1028 | 0.4564 | 0.8274 | 0.9478 | 0.9862 |
| | $m = 15$ | 0.1055 | 0.5876 | 0.9265 | 0.9878 | 0.9988 |
| | $m = 20$ | 0.1032 | 0.6915 | 0.9733 | 0.9974 | 0.9998 |
| | $m = 25$ | 0.0894 | 0.7205 | 0.9797 | 0.9989 | 0.9998 |
| $\alpha = 0.05$ | $m = 5$ | 0.0514 | 0.1699 | 0.4422 | 0.6589 | 0.7971 |
| | $m = 7$ | 0.0512 | 0.2392 | 0.5902 | 0.7998 | 0.9085 |
| | $m = 10$ | 0.0494 | 0.3545 | 0.7611 | 0.9234 | 0.9772 |
| | $m = 15$ | 0.0499 | 0.4812 | 0.8914 | 0.9788 | 0.9973 |
| | $m = 20$ | 0.0504 | 0.5894 | 0.9534 | 0.9955 | 0.9997 |
| | $m = 25$ | 0.0411 | 0.6141 | 0.9633 | 0.9974 | 0.9997 |
| $\alpha = 0.01$ | $m = 5$ | 0.0099 | 0.0692 | 0.2827 | 0.5146 | 0.6858 |
| | $m = 7$ | 0.0096 | 0.1099 | 0.4252 | 0.6785 | 0.8296 |
| | $m = 10$ | 0.0102 | 0.1920 | 0.6107 | 0.8495 | 0.9439 |
| | $m = 15$ | 0.0093 | 0.2889 | 0.7812 | 0.9472 | 0.9901 |
| | $m = 20$ | 0.0094 | 0.3866 | 0.8835 | 0.9857 | 0.9983 |
| | $m = 25$ | 0.0074 | 0.3948 | 0.8931 | 0.9883 | 0.9993 |

**Table 15:** Power for variable window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\theta = 0.5, L = 10000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.1022 | 0.1727 | 0.3381 | 0.4791 | 0.6002 |
| | $m = 7$ | 0.1039 | 0.2225 | 0.4284 | 0.6129 | 0.7282 |
| | $m = 10$ | 0.1071 | 0.2818 | 0.558 | 0.7459 | 0.8579 |
| | $m = 15$ | 0.0970 | 0.3474 | 0.6886 | 0.8704 | 0.9458 |
| | $m = 20$ | 0.1068 | 0.4237 | 0.7935 | 0.9242 | 0.9773 |
| | $m = 25$ | 0.0975 | 0.4707 | 0.8295 | 0.9581 | 0.9902 |
| $\alpha = 0.05$ | $m = 5$ | 0.0487 | 0.1067 | 0.2535 | 0.3933 | 0.5144 |
| | $m = 7$ | 0.0514 | 0.1474 | 0.3445 | 0.5333 | 0.6644 |
| | $m = 10$ | 0.0551 | 0.1923 | 0.4649 | 0.6736 | 0.8018 |
| | $m = 15$ | 0.0491 | 0.2517 | 0.6054 | 0.8194 | 0.9202 |
| | $m = 20$ | 0.0513 | 0.3192 | 0.7127 | 0.8889 | 0.9621 |
| | $m = 25$ | 0.0499 | 0.3609 | 0.7621 | 0.9319 | 0.9798 |
| $\alpha = 0.01$ | $m = 5$ | 0.0076 | 0.0341 | 0.1252 | 0.2403 | 0.3592 |
| | $m = 7$ | 0.0105 | 0.0571 | 0.1972 | 0.3866 | 0.5279 |
| | $m = 10$ | 0.0103 | 0.0748 | 0.2867 | 0.4966 | 0.6692 |
| | $m = 15$ | 0.0104 | 0.1165 | 0.4221 | 0.6822 | 0.8374 |
| | $m = 20$ | 0.0099 | 0.1582 | 0.5317 | 0.7917 | 0.9079 |
| | $m = 25$ | 0.0094 | 0.1772 | 0.5784 | 0.8406 | 0.9429 |

**Table 16:** Power for variable window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\theta = 2, L = 10000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.0974 | 0.4029 | 0.7935 | 0.9438 | 0.9837 |
| | $m = 7$ | 0.1033 | 0.5447 | 0.9111 | 0.9858 | 0.9972 |
| | $m = 10$ | 0.1054 | 0.6954 | 0.9765 | 0.9976 | 0.9998 |
| | $m = 15$ | 0.1032 | 0.8496 | 0.9966 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0936 | 0.914 | 0.9996 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.1004 | 0.9505 | 0.9999 | 1.0000 | 1.0000 |
| $\alpha = 0.05$ | $m = 5$ | 0.0474 | 0.3018 | 0.726 | 0.9163 | 0.9752 |
| | $m = 7$ | 0.0494 | 0.4436 | 0.8724 | 0.9766 | 0.9953 |
| | $m = 10$ | 0.0514 | 0.6 | 0.9609 | 0.9963 | 0.9998 |
| | $m = 15$ | 0.0517 | 0.7851 | 0.9933 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0456 | 0.869 | 0.9992 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0525 | 0.9165 | 0.9998 | 1.0000 | 1.0000 |
| $\alpha = 0.01$ | $m = 5$ | 0.0082 | 0.1492 | 0.5699 | 0.8414 | 0.9458 |
| | $m = 7$ | 0.0081 | 0.2518 | 0.7547 | 0.9417 | 0.9873 |
| | $m = 10$ | 0.0121 | 0.4412 | 0.9206 | 0.9921 | 0.9993 |
| | $m = 15$ | 0.0101 | 0.6174 | 0.982 | 0.9995 | 1.0000 |
| | $m = 20$ | 0.009 | 0.7343 | 0.9975 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0108 | 0.8127 | 0.9988 | 1.0000 | 1.0000 |

**Table 17:** Power for variable window scan statistic (gamma random variables): $N = 100$, true $m = 10$, $\theta = 5, L = 10000$

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.1039 | 0.7771 | 0.994 | 0.9998 | 1.0000 |
| | $m = 7$ | 0.1040 | 0.8961 | 0.9989 | 1.0000 | 1.0000 |
| | $m = 10$ | 0.1061 | 0.9650 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.0989 | 0.9958 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0935 | 0.9995 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.1077 | 0.9999 | 1.0000 | 1.0000 | 1.0000 |
| $\alpha = 0.05$ | $m = 5$ | 0.0520 | 0.7042 | 0.989 | 0.9996 | 1.0000 |
| | $m = 7$ | 0.0509 | 0.8412 | 0.9984 | 1.0000 | 1.0000 |
| | $m = 10$ | 0.0547 | 0.9443 | 0.9999 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.0485 | 0.9926 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0466 | 0.999 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0514 | 0.9998 | 1.0000 | 1.0000 | 1.0000 |
| $\alpha = 0.01$ | $m = 5$ | 0.0107 | 0.5331 | 0.9664 | 0.9984 | 1.0000 |
| | $m = 7$ | 0.0075 | 0.6831 | 0.9947 | 0.9999 | 1.0000 |
| | $m = 10$ | 0.0102 | 0.8677 | 0.9999 | 1.0000 | 1.0000 |
| | $m = 15$ | 0.0089 | 0.9762 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 20$ | 0.0094 | 0.9962 | 1.0000 | 1.0000 | 1.0000 |
| | $m = 25$ | 0.0094 | 0.9991 | 1.0000 | 1.0000 | 1.0000 |

**Table 18:** Power for variable window scan statistic (exponential random variables): $N = 100$, true $m = 10$, L = 10000

| | $m$ | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 2$ | $\beta_1/\beta_0 = 3$ | $\beta_1/\beta_0 = 4$ | $\beta_1/\beta_0 = 5$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | $m = 5$ | 0.1041 | 0.2538 | 0.5349 | 0.7387 | 0.8422 |
| | $m = 7$ | 0.0961 | 0.323 | 0.6753 | 0.8545 | 0.9343 |
| | $m = 10$ | 0.1046 | 0.4362 | 0.8115 | 0.9457 | 0.9859 |
| | $m = 15$ | 0.0985 | 0.5811 | 0.9242 | 0.989 | 0.9979 |
| | $m = 20$ | 0.0996 | 0.6725 | 0.9646 | 0.9983 | 0.9997 |
| | $m = 25$ | 0.0907 | 0.7265 | 0.9803 | 0.999 | 1.0000 |
| $\alpha = 0.05$ | $m = 5$ | 0.0508 | 0.1650 | 0.4361 | 0.6625 | 0.7940 |
| | $m = 7$ | 0.0483 | 0.2349 | 0.5854 | 0.8029 | 0.9077 |
| | $m = 10$ | 0.0528 | 0.3391 | 0.7446 | 0.9204 | 0.9762 |
| | $m = 15$ | 0.0492 | 0.4732 | 0.8837 | 0.9812 | 0.9971 |
| | $m = 20$ | 0.0488 | 0.5743 | 0.942 | 0.9962 | 0.9993 |
| | $m = 25$ | 0.0431 | 0.6222 | 0.9658 | 0.9978 | 1.0000 |
| $\alpha = 0.01$ | $m = 5$ | 0.0081 | 0.0687 | 0.2756 | 0.5112 | 0.6751 |
| | $m = 7$ | 0.0096 | 0.1192 | 0.4301 | 0.689 | 0.8383 |
| | $m = 10$ | 0.0092 | 0.174 | 0.5888 | 0.8394 | 0.9434 |
| | $m = 15$ | 0.0108 | 0.276 | 0.7812 | 0.9529 | 0.9911 |
| | $m = 20$ | 0.0081 | 0.3602 | 0.8648 | 0.9853 | 0.9985 |
| | $m = 25$ | 0.0089 | 0.4165 | 0.9163 | 0.9928 | 0.9998 |

## 3.5 An Application Example: Coal Mine Disasters

To demonstrate an application of the proposed scan statistics, in this section we present an example based on a real data set. This data set is taken from Jarrett (1979), containing 190 observations of the time intervals between the occurrences of coal mine disasters that involved 10 or more deaths, from 15 March 1851 to 22 March 1962. The original data was traced to the *Colliery Year Book and Coal Trades Directory*, available from the National Coal Board in London. The data set has been used extensively in the statistical literature to illustrate a Poisson point process and the exponential distribution (see, e.g., Zhang et al., 2007; Gan, 1998; Maguire et al., 1952).

To obtain a good fit to the exponential distribution, a transformation of $y = \log(x+1)$ is performed, where $x$ is the original time interval in years. Figure 6 presents all the original data points before transformation, where the $x$ axis is the index of observations from 1 to 190, and the $y$ axis is the time intervals in days. Figure 7 shows the quantile-quantile plot of the transformed data with a simulated exponential sample, with the scale parameter estimated from the mean of all observations. A Kolmogorov-Smirnov test is also performed to validate the assumption of exponential distribution, which gives a $p$-value of 0.8464.

We employ scan statistics to detect whether there is a local change in the scale parameter, which indicates a longer average time interval between accidents, i.e. a better and safer working environment for the miners. In this scenario, we do not have

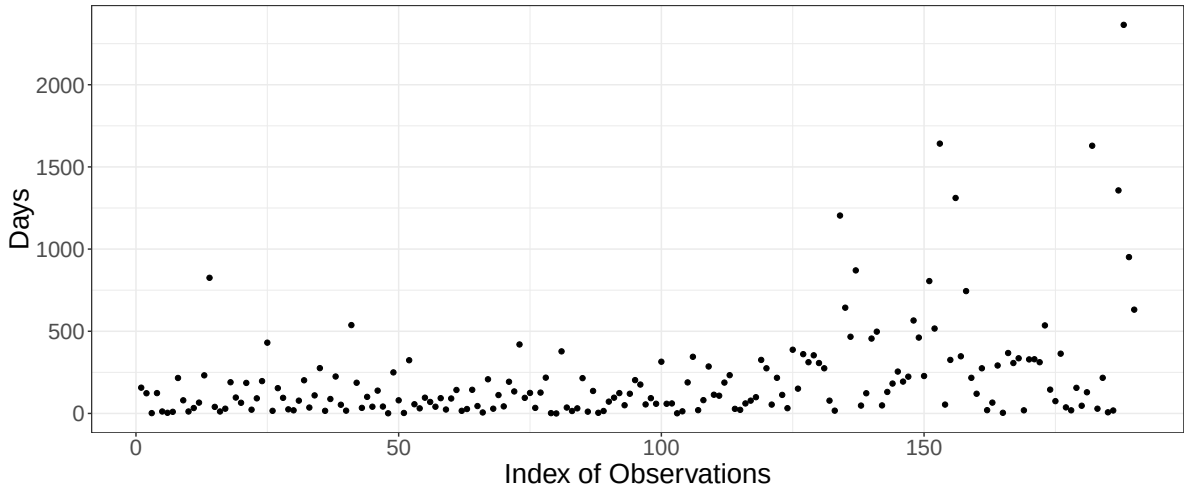


**Figure 6:** Time interval (in days) between coal mine accidents from observations 1 to 190, i.e. from 15 March 1851 to 22 March 1962

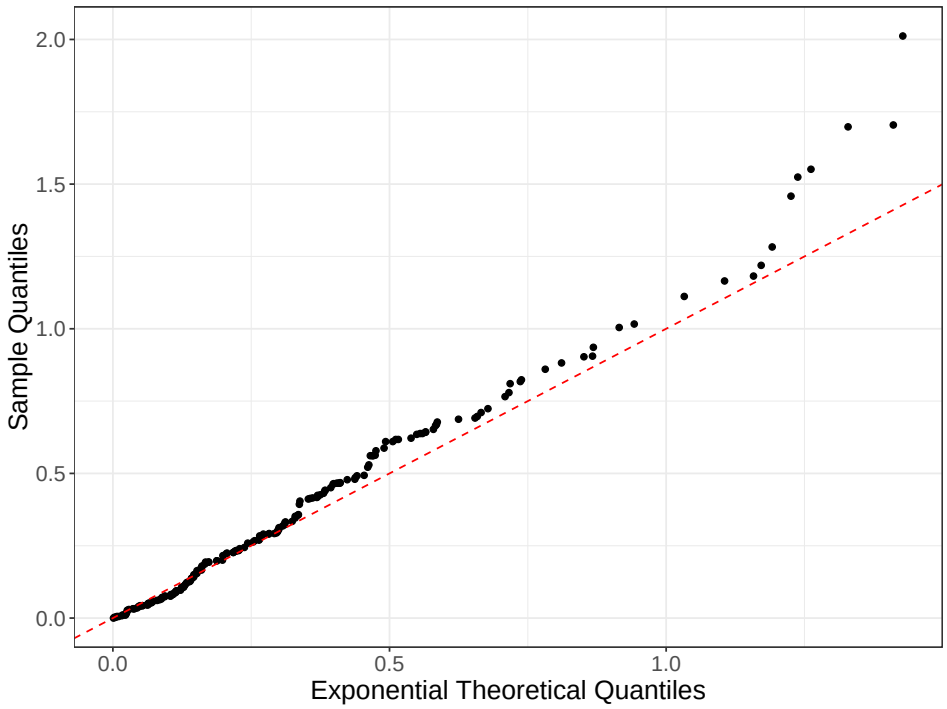


**Figure 7:** Quantile-quantile plot of the transformed data versus an exponential distribution, with the scale parameter $\beta = 213.27$

an estimate for the length of a time period for the occurrence of the local change. Therefore, we cannot decide on a single length of the scanning window. To avoid power loss by using an incorrect window size for the fixed window scan statistic, we need to employ either a multiple window or variable window scan statistic. In this example, there are $N = 190$ observations in total. To implement each of these scan statistic, $M = 500,000$ iterations of simulation are performed.

For the variable window scan statistic, we include all values of $m$ from 5 to 50, and obtained a $p$-value of $3.92 \times 10^{-4}$, with corresponding window length $m = 49$ starting from observation #125. For the multiple window scan statistic we include scanning window lengths of $m \in \{5, 10, 20, 30, 40, 50\}$, yielding a $p$-value $< 4 \times 10^{-6}$ for the $P_{\min}$ statistic. When calculating the minimum $p$-value statistic, although the iteration number is high ($M = 500,000$), the $p$-value obtained for $m = 30, 40, 50$ are all zeros. We include the $p$-value for each conditional fixed window scan statistic and the corresponding starting location of local change in Table 19. The extremely small $p$-values indicate the existence of a upward change of $\beta$, and it is estimated to start around observation #130, which is in the year 1895. In other word, starting from year 1895, the coal miners were able to work in a safer environment in the following 30 to 40 years, but then the accident rate escalated again. The smaller $p$-values for wider window sizes suggest that the level of local change in $\beta$ is small to moderate, therefore it is easier to detect with more observations. The mean of the first 130 data points after transformation is 0.265, and the mean of observation #131 to #170 is 0.613, which is indeed a moderate change. The

**Table 19:** $P$-value of conditional fixed window scan statistics, and estimated local change starting location for $m = 5, 10, 20, 30, 40, 50$

| $m$ | 5 | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|---|
| p-value | $6.4 \times 10^{-3}$ | $1.3 \times 10^{-3}$ | $9.2 \times 10^{-5}$ | 0 | 0 | 0 |
| Starting Obs | 186 | 149 | 134 | 129 | 134 | 125 |

results from the two scan statistics validate each other, and the pattern can be observed from Figure 6 as well. If one knew about the appropriate size for the local change in advance, one could have used the conditional fixed window scan statistic with $m = 50$. This example indicates that the conditional multiple window and variable window scan statistics can be used to detect a moderate local change of the scale parameter for data modeled by exponential distribution.

## 3.6 Concluding Remarks

In this chapter we proposed two different types of scan statistics: the conditional multiple window scan statistic based on the minimum $p$-value method, and the conditional variable window scan statistic based on conditional generalized likelihood ratio test principle. Following these methodologies, we presented algorithms for calculating critical values and the power of these test statistics. Numerical results were presented for selected parameters of the models. Comparisons among conditional fixed, multiple and variable window scan statistics under different parameter settings have been investigated. We have also demonstrated the use of conditional multiple window and variable

window scan statistics with a real data example modeled by exponential distribution, in which both test statistics performed well and can be cross validated.

Based on the numerical results one can conclude that for each of the three scan statistics, if the correct window size is included, the power is increasing as $m$ gets larger, although the extent of growth varies. For example, with $\theta = 1$ and significance level 0.05, for $\beta_1/\beta_0 = 3$ the power is about 0.60 for $m = 7$, but has increased to about 0.80 for $m = 15$, and got close to 1 when $m = 25$. We conclude that the size of the window for detecting a local change is an important factor, regardless of which approach is used.

While the conditional fixed window scan statistic is the most computationally efficient, if the chosen scanning window size of the local change is sufficiently different from the true window size, a major decrease of power could occur, resulting in lower power than all other approaches. To avoid such risk, when a close estimate of the true window size is unobtainable, multiple or variable window scan statistics should be considered.

Recall that while only $\{m_k\}_{k=1}^{K}$ with $2 \leq m_1 < \ldots < m_k \leq N/4$ are considered for the multiple window scan statistic, for the variable window scan statistic, all integers from $m_0$ to $\widetilde{m}_0$, with $3 \leq m_0 < \widetilde{m}_0 \leq N/4$, are considered. In most cases it would be reasonable to assume $K < \widetilde{m}_0 - m_0 + 1$, i.e., the searching scope of multiple window scan statistic is smaller than that of variable window scan statistic, which indicates the former would be more computational efficient. Therefore, in the trade-off between these two methods in practice, especially when the sample size is large, the multiple window scan statistic is recommended as the first choice for testing. Note that one needs to

be cautious in the choice of scanning window sizes $\{m_k\}_{k=1}^{K}$ for multiple window scan statistic, in order to obtain a wide range and a relatively close estimate for the true window size. When the range of window sizes cannot be estimated and the sample size is small to moderate, a variable window scan statistic including a wide range of $m$ values is recommended instead.

# Chapter 4

# Fixed, Multiple and Variable Window Scan Statistics for Two Dimensional Array of Gamma Random Variables

## 4.1 Introduction

In Chapters 2 and 3 we formulated and investigated the performance of one dimensional conditional fixed, multiple and variable window scan statistics to detect a local change in the scale parameter $\beta$ for a sequence of gamma random variables, assuming that the shape parameter $\theta$ is known. The fixed window scan statistic performs best when the true window size $m$ is known, or at least can be reliably estimated, while the multiple and variable window scan statistics are recommended when $m$ is unknown.

In practice, in addition to one dimensional scan statistics, two dimensional scan

statistics are also of interest. Naus (1965, 1966) derived approximations and inequalities for a scan statistic for uniform observations in a two or a higher dimensional unit cube. Chen and Glaz (1996) introduced scan statistics for detecting a local change in the population mean for integer valued observations in a two dimensional region. Alm (1997, 1998) introduced scan statistics for detecting a local change in the intensity of a Poisson process in a rectangular region in two or higher dimensions. Since then two dimensional scan statistics have been investigated for various distributions and applications. In this chapter, similar to Chapters 2 and 3, we discuss the two dimensional conditional fixed, multiple and variable window scan statistics, and evaluate their performance by evaluating the power of these test statistics. Our goal is to detect a local change in the scale parameter $\beta$ for a two dimensional array of gamma random variables, assuming that the shape parameter $\theta$ is known and remains unchanged. Exponential random variables are not investigated in details here, since we state clearly in Chapters 2 and 3 that it is the special case of gamma random variables with $\theta = 1$, and corresponding methodologies can be easily adapted.

This chapter is organized as follows. In Section 4.2 we introduce the two dimensional conditional fixed window scan statistic based on the moving sums of observations in a rectangular window of size $m \times m$ conditional on the sum of all observations. The two dimensional conditional multiple window scan statistic is presented in Section 4.3, based on the minimum $p$-value approach. Section 4.4 investigates the two dimensional variable window scan statistic via the conditional generalized likelihood ratio method. In these

sections we also include algorithms for calculating the critical value of the scan statistics, as well as the power of the tests. In Section 4.5 we present the numerical results from Monte Carlo simulations for all the scan statistics discussed in this chapter, comparing their performance. In Section 4.6 a brief conclusion of the contents of Chapter 4 is presented.

## 4.2 Two Dimensional Fixed Window Scan Statistic

Let $\{X_{k_1,k_2}\}$, $1 \leq k_l \leq N_l, l = 1, 2$ be independent and identically distributed gamma random variables, denoted by $X_{k_1,k_2} \sim \Gamma(\theta, \beta_0)$. Our goal is to detect whether there is a local change of the scale parameter, from $\beta_0$ to $\beta_1$, within a rectangular subregion of observations, in the $N_1 \times N_2$ two dimensional rectangular region. The shape parameter $\theta$ is assumed to be known and constant for all observations. In this chapter we focus on an upward shift in scale parameter. One can easily modify the methodologies presented here to accommodate a local downward or two-sided shift.

The hypotheses we are interested in testing are formulated as following:

$$
\begin{aligned}
&H_0 : X_{k_1,k_2} \sim \Gamma(\theta, \beta_0), \forall k_l = 1, 2, ..., N_l, l = 1, 2; X_{k_1,k_2}\text{'s are independent;} \quad \text{vs.} \\
&H_a : \exists i_l, 1 \leq i_l \leq N_l - m_l + 1, l = 1, 2, \text{such that} \\
&\qquad \forall k_l = i_l, \ldots, i_l + m_l - 1, , X_{k_1,k_2} \sim \Gamma(\theta, \beta_1), \beta_1 > \beta_0; \text{ and} \\
&\qquad \forall k_l = 1, ...i_l - 1, i_l + m_l, ..., N_l, X_{k_1,k_2} \sim \Gamma(\theta, \beta_0); X_{k_1,k_2}\text{'s are independent;}
\end{aligned}
\tag{4.1}
$$

where $\theta$ is known, and $\beta_1$ is unknown. We will discuss both situations where $\beta_0$ is known and $\beta_0$ is unknown.

Denote $2 \leq m_l \leq N_l/4$, $l = 1, 2$, as the prespecified lengths of a two dimensional rectangular sliding window. In an $m_1 \times m_2$ rectangular grid of observed data, the moving sum starting from south west location $\{i_1, i_2\}$ is defined by:

$$Y_{i_1,i_2}(m_1, m_2) = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} X_{k_1,k_2}, \quad 1 \leq i_l \leq N_l - m_l + 1, l = 1, 2. \tag{4.2}$$

The summation of all the observed data, $Y$, is given by:

$$Y = \sum_{k_1=1}^{N_1} \sum_{k_2=1}^{N_2} X_{k_1,k_2}. \tag{4.3}$$

When the scale parameter under the null hypothesis is known, the following *unconditional fixed window scan statistic*, i.e. the maximum moving sum could be used for detecting a local change in scale parameter:

$$S_{m_1,m_2}(N_1, N_2) = \max\{Y_{i_1,i_2}(m_1, m_2) \mid 1 \leq i_l \leq N_l - m_l + 1, l = 1, 2\}. \tag{4.4}$$

Under the null hypothesis, the $X_{k_1,k_2}$'s are i.i.d. gamma random variables. The sequence of moving sums $\{Y_{i_1,i_2}(m_1, m_2), 1 \leq i_l \leq N_l - m_l + 1, l = 1, 2\}$ is stationary and $m$-dependent and it has a special joint multivariate gamma distribution, with identical marginal distributions $\Gamma(m_1 m_2 \theta, \beta_0)$. To simplify the presentation of the results in this

chapter, we assume that $N_1 = N_2 = N$, $m_1 = m_2 = m$, and abbreviate $Y_{i_1,i_2}(m_1, m_2)$, $S_{m_1,m_2}(N_1, N_2)$ to $Y_{i_1,i_2}(m)$ and $S_{m,m}$ respectively.

For $2 \leq m \leq N/4$ and $\infty < t < +\infty$, define:

$$G_{m,t}(N) = P(Y_{i_1,i_2}(m) < t \mid \forall 1 \leq i_1, i_2 \leq N - m + 1) = P(S_{m,m} < t), \tag{4.5}$$

as the cumulative distribution function for $S_{m,m}$. The probability that the scan statistic exceeds level $t$ is given by:

$$P(S_{m,m} \geq t) = 1 - G_{m,t}(N). \tag{4.6}$$

In most applications, however, the scale parameter $\beta_0$ specified in the null hypothesis is unknown. Therefore, the tail probability (4.6) cannot be evaluated. This difficulty can be resolved by conditioning on the sum $Y$ of all observed data, which is the sufficient statistic for $\beta_0$. Under $H_0$, the total sum $Y$ follows a distribution of $\Gamma(N^2\theta, \beta_0)$, and the joint distribution of $X_{1,1}, X_{1,2}, \ldots, X_{N,N}$ conditional on $Y$ can be derived as:

$$\begin{aligned} &f_{X_{1,1},X_{1,2},\ldots,X_{N,N}|Y=y}(x_{1,1}, x_{1,2}, \ldots, x_{N,N}) \\ &= \frac{\Gamma(N^2\theta)}{[\Gamma(\theta)]^{N^2}} \cdot \frac{\left[\prod_{k_1=1}^{N}\prod_{k_2=1}^{N-1} x_{k_1,k_2} \cdot (y - \sum_{k_1=1}^{N}\sum_{k_2=1}^{N-1} x_{k_1,k_2})\right]^{\theta-1}}{y^{N^2\theta-1}}, \end{aligned} \tag{4.7}$$

where $0 < x_{k_1,k_2} < y$, $\forall k_1, k_2 = 1, ..., N$, and $\sum_{k_1=1}^{N}\sum_{k_2=1}^{N} x_{k_1,k_2} = y$. Note that the conditional distribution (4.7) is now free of the scale parameter $\beta_0$. Similarly, under the

null hypothesis, the joint distribution of $X_{i_1,i_2}, ..., X_{i_1+m-1,i_2+m-1},\ 1 \le i_1, i_2 \le N-m+1$ conditional on the partial sum $Y_{i_1,i_2}(m)$ does not depend on $\beta_0$, and can be written as:

$$\begin{aligned} & f_{X_{i_1,i_2},\ldots,X_{i_1+m-1,i_2+m-1}|Y_{i_1,i_2}(m)=\widehat{y}}(x_{i_1,i_2},\ldots,x_{i_1+m-1,i_2+m-1}) \\ &= \frac{\Gamma(m^2\theta)}{[\Gamma(\theta)]^{m^2}} \cdot \frac{\left[\prod_{k_1=i_1}^{i_1+m-1}\prod_{k_2=i_2}^{i_2+m-1} x_{k_1,k_2} \cdot (\widehat{y} - \sum_{k_1=i_1}^{m}\sum_{k_2=i_2}^{m-1} x_{k_1,k_2})\right]^{\theta-1}}{\widehat{y}^{m^2\theta-1}}, \end{aligned} \tag{4.8}$$

where $\widehat{y} = y_{i_1,i_2}(m)$ denotes the specific value of the partial sum random variable $Y_{i_1,i_2}(m)$.

Let $Z_{k_1,k_2} = X_{k_1,k_2}/Y$, then $\sum_{k_1=1}^{N}\sum_{k_2=1}^{N} Z_{k_1,k_2} = 1$. It can be proved that $Z_{1,1}, Z_{1,2}, \ldots, Z_{N,N} \sim \text{Dir}(\vec{\theta}_{N^2})$, where $\vec{\theta}_{N^2} = (\theta, ..., \theta)^{N^2}$ is an $N^2$ dimension vector, $\theta$ being the shape parameter in the gamma distribution of $X_{k_1,k_2}$'s (Devroye, 2013, p. 593). Therefore, we can define the two dimensional *conditional fixed window scan statistic* as below:

$$S^*_{m,m} = \max\{Y^*_{i_1,i_2}(m) \mid 1 \le i_1, i_2 \le N-m+1\}, \tag{4.9}$$

where

$$Y^*_{i_1,i_2}(m) = \sum_{k_1=i_1}^{i_1+m-1}\sum_{k_2=i_2}^{i_2+m-1} Z_{k_1,k_2}, \quad 1 \le i_1, i_2 \le N-m+1. \tag{4.10}$$

We propose to employ this scan statistic to test the hypotheses stated in (4.1). The cumulative distribution of $S^*_{m,m}$ is given by:

$$G^*_{m,t}(N) = P(Y^*_{i_1,i_2}(m) < t \mid \forall i_1, i_2 \le N-m+1) = P(S^*_{m,m} < t). \tag{4.11}$$

If $m$ is known, $H_0$ is rejected when $S^*_{m,m}$ exceeds a certain threshold $t$, where $t$ is determined by $P(S^*_{m,m} > t) = \alpha$, with $\alpha$ being the specified significance level. For a given significance level $\alpha$, the critical value denoted by $p^{(2)}_\alpha$, demonstrates the association between the significance level and the rejection region, i.e.

$$P(S^*_{m,m} > p^{(2)}_\alpha) = \alpha. \tag{4.12}$$

The cumulative distribution function $G^*_{m,t}(N)$ needs to be evaluated before the testing procedure can be implemented. Due to the lack of theoretical results on explicit distribution function or accurate approximations of the multivariate gamma distribution involved, one has to use a Monte Carlo simulation to implement the test procedure. Based on the conditional approach, when performing simulations, if we consider $Z_{k_1,k_2}$'s instead of $X_{k_1,k_2}$'s, $\beta_0$ does not need to be known, which would lead to higher accuracy and efficiency in the applications of scan statistics. In our Monte Carlo simulation, we can do sampling from $\text{Dir}(\vec{\theta}_{N^2})$, where $\theta$ is assumed known, and $X_{k_1,k_2}$ can be calculated by $X_{k_1,k_2} = Z_{k_1,k_2} \cdot Y$ if needed, since $Y$ is considered available from calculating the total sum of all observations (Balakrishnan and Nevzorov, 2004, p.269).

Based on previous discussion on the conditional fixed window scan statistic and sampling method from Dirichlet distribution, to evaluate the performance of the scan statistic $S^*_{m,m}$, we calculate its power for selected parameters in the alternative hypothesis via simulations.

**Algorithm 7:** 2d Fixed Window Scan Statistics: Critical Value for $S^*_{m,m}$

**Result:** Obtain the simulated cumulative probability distribution and the critical value $S^*_{m,m}$ for a given significance level $\alpha$

**for** $r \leftarrow 1$, $R$ **do**
- Draw a sample of $Z_{k_1,k_2}$'s from Dir$(\vec{\theta}_{N^2})$ distribution, $k_1, k_2 = 1, ..., N$;
- Calculate $(N-m+1)^2$ moving sums $Y^*_{i_1,i_2}(m)$;
- Find the maximum moving sum $S^{(r)*}_{m,m}$, and sort them in increasing order as a vector $U^{(r)}_{m,m}$, thus the simulated cumulative probability distribution (4.13) is based on $U^{(r)}_{m,m}$;

**end**

The critical values $p^{(2)}_\alpha$ can be calculated from the simulated $100(1-\alpha)$th percentile of $U^{(r)}_{m,m}$.

**Algorithm 8:** 2d Fixed Window Scan Statistics: Power of Test

**Result:** Test power for alternative hypothesis set by $\beta_1/\beta_0 = 1$, 1.25, 1.5, 1.75, 2

Choose an arbitrary value for the starting position $(i_1, i_2)$ of the local change, $1 \le i_1, i_2 \le N-m+1$, and the total number of simulations $L$;

**for** $l \leftarrow 1$, $L$ **do**
- Generate $X_{i_1,i_2}, ..., X_{i_1+m-1,i_2+m-1} \sim \Gamma(\theta, \beta_1)$, and all other $X_{k_1,k_2}$ outside of this $m \times m$ rectangular subregion from $\Gamma(\theta, \beta_0)$, $1 \le k_1, k_2 \le N$;
- Calculate $Y = \sum_{k_1=1}^{N} \sum_{k_2=1}^{N} X_{k_1,k_2}$, and the maximum moving sum $S^{(l)*}_{m,m}$ following Equation (4.14);
- Compare $S^{(l)*}_{m,m}$ with the simulated cumulative distribution $U^{(r)}_{m,m}$ we obtained in Algorithm 7, and calculate the $p$-value $p^{(l)}_{m,m}$ based on (4.15);

**end**

Calculate the power $\widehat{\eta}^{(2)}_\alpha$ by (4.16).

First we present the algorithm for calculating the critical value of the conditional fixed scan statistic in Algorithm 7. Based on Algorithm 7, the simulated cumulative distribution of the conditional fixed window scan statistic $S^*_{m,m}$ can be derived via:

$$P^*(S^*_{m,m} < t) = r/R, \quad U^{(r)}_{m,m} \leq t < U^{(r+1)}_{m,m}, \tag{4.13}$$

where $r$ is the number of times out of $R$ trials in the simulation that $S^*_{m,m}$ is less than $t$.

Algorithm 8 is developed to calculate the power of the test, i.e. the probability of correctly rejecting the null hypothesis under alternative hypothesis with a specific ratio of $\beta_1/\beta_0$. The simulation studies presented in Section 4.5 include results with $\beta_1 \in \{1, 1.25, 1.5, 1.75, 2\}$ and $\beta_0 = 1$. By conditioning on the total sum of all observations $Y$, we can employ the conditional fixed window scan statistic (2.9), which distribution does not depend on any unknown parameter. The maximum moving sum in Step 4 is defined as:

$$S^{(l)*}_{m,m} = \max\{Y^*_{i_1,i_2}(m) \mid 1 \leq i_1, i_2 \leq N - m + 1\}. \tag{4.14}$$

The $p$-value is calculated by:

$$p^{(l)}_{m,m} = P^*(S^{(l)*}_{m,m} < t) = r/R, \quad U^{(r)}_{m,m} \leq t < U^{(r+1)}_{m,m}, \tag{4.15}$$

where $r$ is the number of times out of $R$ trials in the simulation that $S^{(l)*}_{m,m}$ is less than $t$.

Power calculation is based on:

$$\widehat{\eta}_{\alpha}^{(2)} = \frac{\#\{p_{m,m}^{(l)} < \alpha, l = 1, ..., L\}}{L}. \tag{4.16}$$

Simulation study based on these algorithms is performed and the results are shown in Section 4.5.

## 4.3 Two Dimensional Multiple Window Scan Statistic

The application of fixed window scan statistic in practice is limited, since it only tests for one window size, indicating that it is best used in the case that the true size is known for the $m \times m$ rectangular region where the local change of $\beta$ occurs. Using an incorrect size of the moving window, on the other hand, will results in loss of power. To handle the case that $m$ is unknown, one approach is to employ *conditional multiple window scan statistic* (Zhao and Glaz, 2016). One can take into consideration simultaneously a sequence of $K$ window sizes $\{m_k \times m_k\}_{k=1}^{K}$, where $2 \leq m_1 < \ldots < m_K \leq N/4$ are chosen in advance by the experimenter.

We are interested in testing the hypotheses problem stated in (4.1). The conditional fixed window scan statistics $S_{m_k,m_k}^{*}$ for $\{m_k\}_{k=1}^{K}$ then can be calculated following Equation (4.9). Let $t_k$ be the observed value of $S_{m_k,m_k}^{*}$, the associated $p$-value $p_{m_k}^{(2)}$ is defined

as:

$$p_{m_k}^{(2)} = P(S_{m_k,m_k}^{*} > t_k \mid H_0). \tag{4.17}$$

To test the hypothesis in (4.1), we propose to use the minimum $p$-value statistic, denoted as $P_{\min}^{(2)}$, as the test statistic:

$$P_{\min}^{(2)} = \min\{p_{m_k}^{(2)}; 1 \le k \le K\}. \tag{4.18}$$

The null hypothesis is rejected if the observed value of $P_{\min}^{(2)}$ falls below a critical value $p_{\alpha}^{K(2)}$ based on a prespecified significance level $\alpha$:

$$P(P_{\min}^{(2)} < p_{\alpha}^{K(2)}) = \alpha. \tag{4.19}$$

The implementation of the $P_{\min}$ statistic can be carried out by a Monte Carlo simulation. Algorithm 9 can be used to find the critical value $p_{\alpha}^{K(2)}$.

The simulated cumulative distribution of $S_{m_k,m_k}^{*}$ can be derived via:

$$P^{*}(S_{m_k,m_k}^{*} < t_k) = r/R, \quad U_{m_k}^{(r)} \le t_k < U_{m_k}^{(r+1)}, \tag{4.20}$$

where $r$ is the number of times out of $R$ trials in the simulation that $S_{m_k,m_k}^{*}$ is less than $t_k$. The $p$-value $p_{m_k}^{(2)(s)}$ is defined as:

$$p_{m_k}^{(2)(s)} = P^{*}(S_{m_k,m_k}^{(s)*} > t_k) = r/R, \quad U_{m_k}^{(r)} \le t_k < U_{m_k}^{(r+1)}, \tag{4.21}$$

**Algorithm 9:** 2d Multiple Window Scan Statistics: Critical Value for $P_{\min}^{(2)}$

**Result:** Obtain the simulated cumulative probability distribution and the critical value $p_\alpha^{K(2)}$ for $P_{\min}^{(2)}$ with a given significance level $\alpha$

: **for** $r \leftarrow 1$, $R$ **do**
  Draw a sample of $Z_{k_1,k_2}$'s from Dir($\vec{\theta}_{N^2}$) distribution, $k_1, k_2 = 1, \ldots, N$;
  **for** $m_k \in \{m_k \mid k = 1, \ldots, K\}$ **do**
    Calculate $(N - m_k + 1)^2$ moving sums $Y^*_{i_1,i_2}(m_k)$;
    Find the maximum moving sum $S^{(r)*}_{m_k,m_k}$, and sort them in increasing order as a vector $U^{(r)}_{m_k}$, thus the simulated cumulative distribution function (3.4) is based on $U^{(r)}_{m_k}$.
  **end**
**end**
**for** $s \leftarrow 1, S$ **do**
  Simulate a sample of $Z_{k_1,k_2}$'s from Dir($\vec{\theta}_{N^2}$) distribution, $k_1, k_2 = 1, \ldots, N$;
  Calculate $S^{(s)*}_{m_k,m_k}$ for each $m_k$ in the set $\{m_k \mid k = 1, \ldots, K\}$;
  **for** $m_k \in \{m_k \mid k = 1, \ldots, K\}$ **do**
    obtain the $p$-value $p^{(2)(s)}_{m_k}$;
  **end**
  Compute the minimum $p$-value $P^{(2)(s)}_{\min}$;
**end**
Sort $\{P^{(2)(s)}_{\min} \mid s = 1, \ldots, S\}$ in ascending order as a vector $Q^{(2)(s)}$, and the critical values $p_\alpha^{K(2)}$ can be calculated from the simulated $100(1-\alpha)$th percentile of $Q^{(2)(s)}$.

where $r$ is the number of times out of $R$ trials in the simulation that $S^{(s)*}_{m_k,m_k}$ is larger than $t_k$. Test statistic $P^{(s)}_{\min}$ is given by:

$$P^{(2)(s)}_{\min} = \min\{p^{(2)(s)}_{m_k}; 1 \leq k \leq K\}. \tag{4.22}$$

Algorithm 10 is demonstrated below to calculate the power of test for specific alternative hypothesis settings. The simulation studies presented in Section 4.5 include results with $\beta_1 \in \{1, 1.25, 1.5, 1.75, 2\}$ and $\beta_0 = 1$. The $p$-value in step 9, $p^{(2)(l)}$, can be calculated from the simulated cumulative distribution from Algorithm 9:

$$p^{(2)(l)} = P^*(P^{(2)(l)}_{\min} < t) = s/S, \quad Q^{(2)(s)} \leq t < Q^{(2)(s+1)}, \tag{4.23}$$

where $s$ is the number of times out of $S$ trials in the simulation that $p^{(2)(l)}$ is less than $t$. And the power of the test statistic $P_{\min}$, $\widehat{\eta}^{(2)}_{\alpha}$ is defined as:

$$\widehat{\eta}^{(2)}_{\alpha} = \frac{\#\{p^{(2)(l)} < \alpha, l = 1, ..., L\}}{L}. \tag{4.24}$$

**Algorithm 10:** 2d Multiple Window Scan Statistics: Power of Test

**Result:** Power of the test for specific alternative hypotheses

Choose an arbitrary value for the starting position $(i_1, i_2)$ of the local change, $1 \leq i_1, i_2 \leq N - m + 1$, and the total number of simulations $L$;
**for** $l \leftarrow 1$, $L$ **do**
    Generate $X_1, ..., X_{i_1-1,i_2-1} \sim \Gamma(\theta, \beta_0)$, $X_{i_1,i_2}, ..., X_{i_1+m-1,i_2+m-1} \sim \Gamma(\theta, \beta_1)$, and $X_{i_1+m,i_2+m}, ..., X_{N,N} \sim \Gamma(\theta, \beta_0)$;
    **for** $m_k \in \{m_k | k = 1, ..., K\}$ **do**
        Calculate $S^{(l)*}_{m_k,m_k}$;
        Compute the associated $p$-value $p^{(2)(l)}_{m_k}$ for $S^{(l)*}_{m_k,m_k}$;
    **end**
    Calculate the minimum $p$-value statistic $P^{(2)(l)}_{\min}$;
    Compare $P^{(2)(l)}_{\min}$ to the simulated cumulative distribution $Q^{(2)(s)}$ we obtained from Algorithm 3, and calculate the $p$-value for this single iteration $p^{(2)(l)}$.
**end**
Obtain the power $\widehat{\eta}^{(2)}_{\alpha}$.

## 4.4 Two Dimensional Variable Window Scan Statistic

When the size of the $m \times m$ rectangular region where a local change of $\beta$ occurs is unknown, one can employ an alternative approach to a multiple window scan statistic: a *variable window scan statistic* based on the generalized likelihood ratio test principle (Kulldorff, 1997; Nagawalla, 1996). We employ a conditional generalized likelihood ratio test (cGLRT), conditioning on the total sum of the data sequence $Y = \sum_{k_1=1}^{N} \sum_{k_2=1}^{N} X_{k_1,k_2}$, as well as the partial sum $Y_{i_1,i_2}(m)$. Assume that the true size $m$ of the $m \times m$ region of the local change is between $m_0$ and $\widetilde{m}_0$, with $3 \leq m_0 \leq m \leq \widetilde{m}_0 \leq N/4$. For our hypothesis testing problem stated in (4.1), the generalized

ratio test should reject $H_0$ in favor of $H_1$ for large values of:

$$\Lambda^{(2)} = \frac{\sup_{\Theta_1} f(x_{1,1}, ..., x_{N,N}|y, y_{i_1,i_2}(m))}{\sup_{\Theta_0} f(x_{1,1}, ..., x_{N,N}|y, y_{i_1,i_2}(m))}, \tag{4.25}$$

where $f(x_{1,1}, ..., x_{N,N}|y, y_{i_1,i_2}(m))$ is the joint distribution of $X_{1,1}, \ldots, X_{N,N}$ conditional on the total sum $Y$ and the partial sum $Y_{i_1,i_2}(m)$, $\Theta_0$ and $\Theta_1$ denote the respective parameter spaces under $H_0$ and $H_1$. It can be simplified to:

$$\begin{aligned}
\Lambda^{(2)} =& \Lambda^{(2)}(i_1, i_2, m \mid y, y_{i_1,i_2}) \\
=& \sup_{\Theta_1} \left\{ \frac{\Gamma(m^2\theta)\Gamma[(N^2 - m^2)\theta]}{\Gamma(N^2\theta)} \cdot \frac{y^{N^2\theta-1}}{y_{i_1,i_2}^{m^2\theta-1}(y - y_{i_1,i_2})^{(N^2-m^2)\theta-1}} \right\} \\
=& \sup_{i_1,i_2,m} \left\{ \frac{\Gamma(m^2\theta)\Gamma[(N^2 - m^2)\theta]}{\Gamma(N^2\theta)} \cdot \frac{y}{(y_{i_1,i_2}/y)^{m^2\theta-1}(1 - y_{i_1,i_2}/y)^{(N^2-m^2)\theta-1}} \right\},
\end{aligned} \tag{4.26}$$

where we write $y_{i_1,i_2} = y_{i_1,i_2}(m)$ for simplification. Define a function

$$h(v) = v^{1-m^2\theta}(1 - v)^{1-(N^2-m^2)\theta},$$

where $v = y_{i_1,i_2}/y, 0 < v < 1$. (4.26) can be further simplified to:

$$\Lambda^{(2)} = \sup_{i_1,i_2,m} \left\{ \frac{\Gamma(m^2\theta)\Gamma[(N^2 - m^2)\theta]y}{\Gamma(N^2\theta)} \cdot h(v) \right\}. \tag{4.27}$$

Since the sum of all observations $y$ is also known, and we assume that the shape parameter $\theta$ is known, given $m$ is fixed, the first part of Equation (4.27) is a constant. To find

the maximum value of the likelihood function, we only need to find $\max\{h(v)\}$.

Because $h(v)$ is a function of $v = y_{i_1,i_2}/y = y_{i_1,i_2}(m)/y$, it depends on $(i_1, i_2)$ and $m$. For a given but arbitrary $m$, $h(v)$ is a convex function on $v$. In this case, the maximum value of the function should be obtained at either the smallest or the largest value of $v$. With the alternative hypothesis being $\beta_1 > \beta_0$, the expectation of the observations within the rectangular sub-region of local change of $\theta$ would be larger than that of the remaining $X_{k_1,k_2}$'s outside of the sub-region. Therefore the smallest value of $v$, i.e. a small ratio between the partial sum and the total sum, would not lead to a conclusion in favor of $H_a$. We only need to obtain $\max\{y_{i_1,i_2}/y\}$, which is exactly the conditional fixed window scan statistic $S^*_{m,m}$ defined in Section 4.2.

For a given $m$, we need to find $S^*_{m,m}$ in the sample and calculate the corresponding likelihood ratio, denoted by $L^*_{m,m}$:

$$L^*_{m,m} = \frac{\Gamma(m^2\theta)\Gamma[(N^2-m^2)\theta]}{\Gamma(N^2\theta)} \cdot \frac{y}{(S^*_{m,m})^{m^2\theta-1}(1-S^*_{m,m})^{(N^2-m^2)\theta-1}}. \tag{4.28}$$

The next step is to compute $L^*_{m,m}$ for all $m$ values between $m_0$ and $\widetilde{m}_0$, which in the most complete case would be all the integers between 3 and $N/4$. We can obtain the variable window scan statistic $\Lambda^{(2)*}$ as:

$$\Lambda^{(2)*} = \max\{L^*_{m,m} \mid m_0 \le m \le \widetilde{m}_0\}. \tag{4.29}$$

We denote the corresponding $m$ value as $m^*$, and record the location $(i_1^*(m^*), i_2^*(m^*))$ of the maximum moving sum $S^*_{m^*,m^*}$, for which we believe the $m^* \times m^*$ region starting at location $(i_1(m^*), i_2(m^*))$ is the most likely where the local change of $\beta_0$ occurs.

In Algorithm 11 we summarize how to find the variable window scan statistic. Monte Carlo simulations is used to find the simulated cumulative distribution of the test statistic $\Lambda^{(2)*}$, along with the critical value $p_\alpha^{\Lambda(2)}$, which we present in Algorithm 12.

The simulated cumulative distribution of the variable window scan statistic $\Lambda^{(2)*}$ can be derived via:

$$P^*(\Lambda^{(2)*} < t) = r/R, \quad \Phi^{(r)} \leq t < \Phi^{(r+1)}, \tag{4.30}$$

where $r$ is the number of times out of $R$ trials in the simulation that $\Lambda^{(2)*}$ is less than $t$.

The power of the variable window scan statistic can also be calculated in a similar way to the fixed window scan statistic, via Monte Carlo simulation, as presented in Algorithm 8.

**Algorithm 11:** 2d Variable Window Scan Statistic: Maximum Likelihood Ratio

**Result:** The variable window scan statistic, i.e. maximum likelihood ratio $\Lambda^{(2)*}$

**for** $m \leftarrow m_0$, $\widetilde{m}_0$ **do**

  Compute moving sums $Y^*_{i_1,i_2}(m)$, $\forall i_1, i_2 = 1, \ldots, N - m + 1$;

  Find the maximum moving sum $S^*_{m,m}$, and record the corresponding starting location $(i_1(m), i_2(m))$;

  Calculate $L^*_{m,m}$ by Equation (4.28);

**end**

Find $\Lambda^{(2)*}$ as defined in (4.29), and the corresponding $m^*$ and $(i_1(m^*), i_2(m^*))$.

**Algorithm 12:** 2d Variable Window Scan Statistic: Critical Value of $\Lambda^{(2)*}$

**Result:** The simulated cumulative distribution and the critical value of $\Lambda^{(2)*}$

**for** $r \leftarrow 1$, $R$ **do**

  Simulate a sample of $Z_{k_1,k_2}$'s that follow $\mathrm{Dir}(\vec{\theta}_{N^2})$ distribution, $k_1, k_2 = 1, ..., N$;

  Calculate the variable window scan statistic $\Lambda^{(2)(r)*}$ as defined in (4.29), and sort them in increasing order as a vector $\Phi^{(r)}$, thus the simulated cumulative distribution function (4.30) is based on $\Phi^{(r)}$;

**end**

The critical values $p_\alpha^{\Lambda(2)}$ can be calculated from the simulated $100(1-\alpha)$th percentile of $\Phi^{(r)}$.

## 4.5 Numerical Results

In Sections 4.2, 4.3 and 4.4, we derived two dimensional conditional fixed window, multiple window, and variable window scan statistics. We now present numerical results for the power of these test statistics and compare their performance for selected parameters of the alternative hypothesis. In Figure 8, we present the power for multiple window and variable window scan statistics, as well as that of the conditional fixed window scan statistic. These power comparisons are all based on alternative hypothesis with samples generated with a local change in an $m \times m$ rectangular sub-region with $m = 10$, within an $N \times N$ rectangular region with $N = 100$, and under different ratios of $\beta_1/\beta_0 \in \{1, 1.25, 1.5, 1.75, 2\}$. Power is calculated from $L = 10,000$ Monte Carlo simulations, presented with significance levels $\alpha = 0.10$, 0.05 and 0.01. The corresponding power values are recorded in Tables 21 to 24. In Tables 21, 22 and 23, we present results for gamma random variables, with $\theta = 0.5$, 2, and 5, respectively. In Table 24 numerical results are presented for observations from the exponential distribution.

Abbreviations used in the Tables below, are summarized in Table 20 for reference.

**Table 20:** Approach Abbreviations

| Approach | Scan Statistic | True $m$ | $m$ tested | Power Rank |
|---|---|---|---|---|
| FW(T) | Fixed window scan statistic | 10 | 10 | 1 |
| FW(F) | Fixed window scan statistic | 10 | 5 | 5 |
| MW(I) | Multiple window scan statistic | 10 | 5, 10, 20 | 2 (tie) |
| MW(O) | Multiple window scan statistic | 10 | 15, 20, 25 | 4 |
| VW | Variable window scan statistic | 10 | 3 - 25 | 2 (tie) |

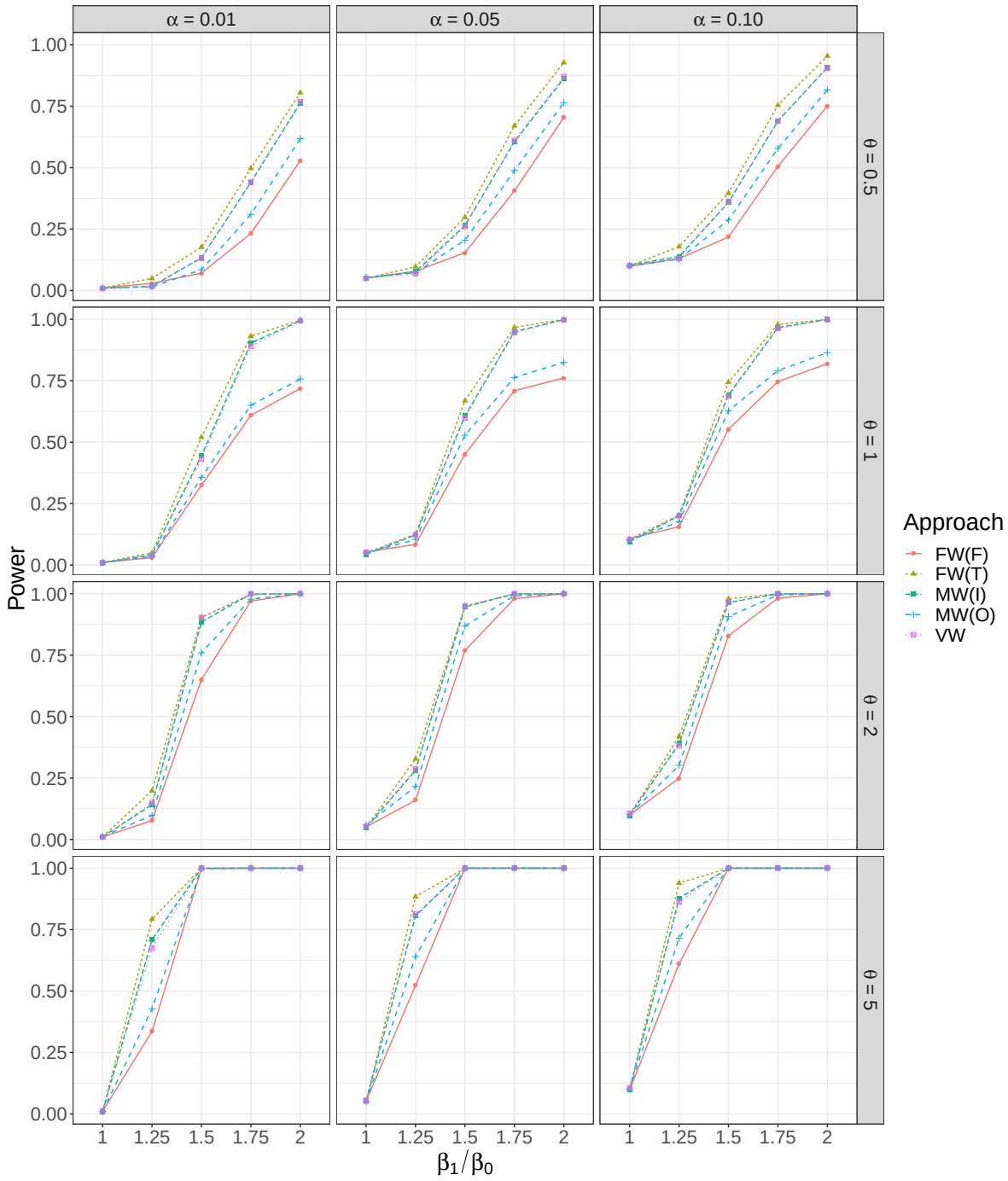


**Figure 8:** Power comparison for 2d scan statistics gamma random variables: $N = 100$, true $m = 10$, $\theta = 0.5, 1, 2$ and 5

For the different approaches, "FW(T)" stands for fixed window scan statistic with the correct window size of local change ($m = 10$), "FW(F)" is for fixed window scan statistic with a wrong window size ($m = 5$), "MW(I)" indicates multiple window scan statistic with the window sizes tested including $m = 10$ ($m_k = 5, 10$ and 20), "MW(O)" means multiple window scan statistic with the true window size falling outside of the tested window lengths ($m_k = 15$, 20, and 25 ), and "VW" for variable window scan statistic with a testing range of $m$ from 3 to 25.

In Figure 2, the four rows of plots demonstrate power results for gamma random variables with shape parameter $\theta = 0.5$, 1, 2 and 5. Note that gamma random variables with $\theta = 1$ are essentially exponential random variables, and we are presenting in the current order only for the convenience of showing the trend with the increasing $\theta$. The three columns of plots correspond to significance level $\alpha = 0.01$, 0.05 and 0.10 from right to left.

When the local shift of $\beta_1/\beta_0$ in the alternative hypothesis is very small, none of the proposed scan statistics show high power in detecting the local change. But as the local shift of $\beta_1/\beta_0$ becomes larger, all of the three approaches perform well except for the fixed window scan statistics using an incorrect window size (far away from true size). The power of these approaches can be ranked from in descending order: FW(T) $>$ MW(I) $\approx$ VW $>$ MW(O) $>$ FW(F). The rankings are also presented in Table 20. The fixed window scan statistic using the exact true window size shows the highest power, as well as the best computational efficiency. Multiple and variable window scan statistics

that cover the true window size show slightly lower power but still both perform well, and the difference between these two is not substantial, usually around 0.02 or smaller. Multiple window scan statistic with the true window size falling outside of the tested set show further lower power than these three, and the powers of fixed window scan statistic using an incorrect window size are the lowest.

Taking the case of $\theta = 1$ in Table 24 as an example, we can see with ratio of $\beta_1/\beta_0 = 1.75$, which is not a substantial shift, even with a significance level of $\alpha = 0.05$, the powers of FW(T), MW(I) and VW have reached around 0.95. In other words, 95% of the time such shift in the scale parameter would be detected with these approaches, indicating that our methods perform excellent in a medium shift of $\beta_1/\beta_0$.

For all the tests, given the same $\beta_1/\beta_0$ and significance level $\alpha$, simulations with larger shape parameter $\theta$ yield higher power, possibly due to $E(X_i) = \theta\beta$, indicating that larger $\theta$ would lead to a larger local shift for observations inside the window.

Comparing to the power results of one dimensional scan statistics in Section 3.4, power of test increase faster as the ratio of $\beta_1/\beta_0$ becomes larger, reaching 1 at a relatively lower ratio. The differences between different approaches also become larger. The power decrease from FW(T) to FW(F) is more considerable now, indicating that in two dimensional case, an incorrect choice of the scanning window size would be more influential. This is expected because there are much more observations with one more dimension. We also run simulations for different true rectangular region sizes with the same $\{m_k\}_{k=1}^{K}$ set for multiple window scan statistic, and the results show that the

power increases as the true rectangular region size increases, which is natural since a local change of $\beta$ would be easier to detect if the corresponding area is larger. Similar conclusions remain valid for the variable window scan statistic.

**Table 21:** Power comparison for 2d gamma random variables: $N = 100$, true $m = 10$, $\theta = 0.5, L = 10000$

| | Approach | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 1.25$ | $\beta_1/\beta_0 = 1.5$ | $\beta_1/\beta_0 = 1.75$ | $\beta_1/\beta_0 = 2$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | FW(T) | 0.1001 | 0.1782 | 0.3962 | 0.7544 | 0.9541 |
| | FW(F) | 0.0982 | 0.1303 | 0.2177 | 0.5050 | 0.7509 |
| | MW(I) | 0.1033 | 0.1385 | 0.3583 | 0.6899 | 0.9079 |
| | MW(O) | 0.1007 | 0.1274 | 0.2867 | 0.5784 | 0.8162 |
| | VW | 0.1017 | 0.1303 | 0.3621 | 0.6896 | 0.9053 |
| $\alpha = 0.05$ | FW(T) | 0.0492 | 0.0970 | 0.2980 | 0.6699 | 0.9277 |
| | FW(F) | 0.0511 | 0.0782 | 0.1544 | 0.4050 | 0.7042 |
| | MW(I) | 0.0505 | 0.0766 | 0.2648 | 0.6051 | 0.8631 |
| | MW(O) | 0.0527 | 0.0725 | 0.2044 | 0.4884 | 0.7643 |
| | VW | 0.0494 | 0.0685 | 0.2600 | 0.6109 | 0.8706 |
| $\alpha = 0.01$ | FW(T) | 0.0102 | 0.0493 | 0.1766 | 0.4972 | 0.8052 |
| | FW(F) | 0.0112 | 0.0290 | 0.0711 | 0.2322 | 0.5287 |
| | MW(I) | 0.0084 | 0.0180 | 0.1330 | 0.4406 | 0.7631 |
| | MW(O) | 0.0105 | 0.0151 | 0.0847 | 0.3096 | 0.6181 |
| | VW | 0.0086 | 0.0168 | 0.1310 | 0.4408 | 0.7687 |

**Table 22:** Power comparison for 2d gamma random variables: $N = 100$, true $m = 10$, $\theta = 2, L = 10000$

| | Approach | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 1.25$ | $\beta_1/\beta_0 = 1.5$ | $\beta_1/\beta_0 = 1.75$ | $\beta_1/\beta_0 = 2$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | FW(T) | 0.0979 | 0.4189 | 0.9789 | 1.0000 | 1.0000 |
| | FW(F) | 0.0996 | 0.2493 | 0.8287 | 0.9822 | 1.0000 |
| | MW(I) | 0.0981 | 0.3901 | 0.9642 | 1.0000 | 1.0000 |
| | MW(O) | 0.1017 | 0.3018 | 0.9071 | 0.9950 | 1.0000 |
| | VW | 0.1065 | 0.3811 | 0.9667 | 1.0000 | 1.0000 |
| $\alpha = 0.05$ | FW(T) | 0.0492 | 0.3280 | 0.9503 | 0.9978 | 1.0000 |
| | FW(F) | 0.0511 | 0.1594 | 0.7691 | 0.9801 | 1.0000 |
| | MW(I) | 0.0479 | 0.2809 | 0.9459 | 1.0000 | 1.0000 |
| | MW(O) | 0.0584 | 0.2156 | 0.8695 | 0.9913 | 1.0000 |
| | VW | 0.0534 | 0.2861 | 0.9500 | 1.0000 | 1.0000 |
| $\alpha = 0.01$ | FW(T) | 0.0102 | 0.1982 | 0.9037 | 0.9965 | 1.0000 |
| | FW(F) | 0.0103 | 0.0766 | 0.6506 | 0.9702 | 1.0000 |
| | MW(I) | 0.0107 | 0.1410 | 0.8850 | 0.9990 | 1.0000 |
| | MW(O) | 0.0122 | 0.0978 | 0.7597 | 0.9781 | 1.0000 |
| | VW | 0.0098 | 0.1493 | 0.9030 | 0.9998 | 1.0000 |

**Table 23:** Power comparison for 2d gamma random variables: $N = 100$, true $m = 10$, $\theta = 5, L = 10000$

| | Approach | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 1.25$ | $\beta_1/\beta_0 = 1.5$ | $\beta_1/\beta_0 = 1.75$ | $\beta_1/\beta_0 = 2$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | FW(T) | 0.1002 | 0.9394 | 1.0000 | 1.0000 | 1.0000 |
| | FW(F) | 0.0960 | 0.6099 | 1.0000 | 1.0000 | 1.0000 |
| | MW(I) | 0.0997 | 0.8769 | 1.0000 | 1.0000 | 1.0000 |
| | MW(O) | 0.1010 | 0.7154 | 1.0000 | 1.0000 | 1.0000 |
| | VW | 0.1064 | 0.8622 | 1.0000 | 1.0000 | 1.0000 |
| $\alpha = 0.05$ | FW(T) | 0.0489 | 0.8840 | 1.0000 | 1.0000 | 1.0000 |
| | FW(F) | 0.0513 | 0.5230 | 1.0000 | 1.0000 | 1.0000 |
| | MW(I) | 0.0521 | 0.8058 | 1.0000 | 1.0000 | 1.0000 |
| | MW(O) | 0.0497 | 0.6407 | 1.0000 | 1.0000 | 1.0000 |
| | VW | 0.0568 | 0.8128 | 1.0000 | 1.0000 | 1.0000 |
| $\alpha = 0.01$ | FW(T) | 0.0101 | 0.7925 | 1.0000 | 1.0000 | 1.0000 |
| | FW(F) | 0.0109 | 0.3355 | 0.9990 | 1.0000 | 1.0000 |
| | MW(I) | 0.0093 | 0.7088 | 1.0000 | 1.0000 | 1.0000 |
| | MW(O) | 0.0096 | 0.4280 | 0.9986 | 1.0000 | 1.0000 |
| | VW | 0.0128 | 0.6749 | 1.0000 | 1.0000 | 1.0000 |

**Table 24:** Power comparison for 2d exponential random variables: $N = 100$, true $m = 10$, $L = 10000$

| | Approach | $\beta_1/\beta_0 = 1$ | $\beta_1/\beta_0 = 1.25$ | $\beta_1/\beta_0 = 1.5$ | $\beta_1/\beta_0 = 1.75$ | $\beta_1/\beta_0 = 2$ |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | FW(T) | 0.1041 | 0.2031 | 0.7439 | 0.9785 | 0.9996 |
| | FW(F) | 0.1080 | 0.1554 | 0.5525 | 0.7465 | 0.8190 |
| | MW(I) | 0.0954 | 0.2004 | 0.6929 | 0.9664 | 0.9992 |
| | MW(O) | 0.1011 | 0.1758 | 0.6275 | 0.7909 | 0.8641 |
| | VW | 0.1032 | 0.2004 | 0.6860 | 0.9628 | 0.9982 |
| $\alpha = 0.05$ | FW(T) | 0.0492 | 0.1255 | 0.6684 | 0.9670 | 0.9992 |
| | FW(F) | 0.0534 | 0.0836 | 0.4512 | 0.7090 | 0.7611 |
| | MW(I) | 0.0436 | 0.1223 | 0.6066 | 0.9500 | 0.9981 |
| | MW(O) | 0.0490 | 0.1037 | 0.5283 | 0.7629 | 0.8246 |
| | VW | 0.0506 | 0.1229 | 0.5984 | 0.9460 | 0.9975 |
| $\alpha = 0.01$ | FW(T) | 0.0102 | 0.0493 | 0.5197 | 0.9314 | 0.9966 |
| | FW(F) | 0.0106 | 0.0301 | 0.3248 | 0.6089 | 0.7184 |
| | MW(I) | 0.0085 | 0.0421 | 0.4443 | 0.9032 | 0.9937 |
| | MW(O) | 0.0105 | 0.0345 | 0.3567 | 0.6502 | 0.7566 |
| | VW | 0.0098 | 0.0405 | 0.4303 | 0.8905 | 0.9936 |

## 4.6 Concluding Remarks

In this chapter we presented two dimensional fixed window, multiple window, and variable window scan statistics, for detecting a local change of the scale parameter $\beta$ occurring in a rectangular sub-region $m \times m$, within an $N \times N$ rectangular region of observations from a gamma distribution. We assumed that the shape parameter $\theta$ is known and remains constant for all $N \times N$ observations. Algorithms for calculating the critical values and power of the test statistics are developed in each section for different type of scan statistics. Based on these algorithms Monte Carlo simulations were performed, and numerical results were presented and discussed.

From the computational efficiency perspective, the fixed window scan statistic (FW) is most efficient, followed by the multiple window scan statistic (MW). The variable window scan statistic (VW) takes the longest time to implement and compute its power. This is because FW only scans one $m$ value, and in our simulation MW scans three, and VW scans 23 values, i.e. including all possible $m$ values up to $N/4$. The selection of the set $\{m_k\}_{k=1}^{K}$ is usually based on the experimenter's experience and knowledge of the possible size for the local change window, which is not designed to select all possible value such as VM. Therefore MW is generally more computational efficient than VM. The consideration on computational efficiency can be crucial when $N$ is large, especially in higher dimensional data, such as in applications in genetics or epidemiology.

Based on our comparison in Section 4.5, when the size of the local change window can

be relatively accurately estimated, fixed window scan statistic is recommended due to its superior power and best computational speed. On the other hand, if the tested scanning window size is far from the true size of local change, there is a major decrease in power for fixed window scan statistic. In this case, multiple window or variable window scan statistics should be considered. In choosing between these two, if one is confident that the true window size would fall inside the set $\{m_k\}_{k=1}^{K}$, multiple window scan statistic is appropriate; otherwise a variable window scan statistic including a wide range of $m$ values is recommended.

# Chapter 5

# Summary

## 5.1 Conclusion

In this dissertation, we have investigated the use of scan statistics for detecting a local change in the scale parameter $\beta$ of observations from a gamma distribution in one and two-dimensional regions, when the shape parameter is known and constant. We have focused on an upward shift of $\beta$, and the methodologies can also be accommodated to detect a downward or two-sided shift. When the true window size $m$ where the local shift occurs is known, and the scale parameter in the null hypothesis is unknown, one can employ a conditional fixed window scan statistic. If the true window size for a local change in the scale parameter is unknown, conditional multiple window scan statistic via the minimum $p$-value approach and variable window scan statistic using the conditional generalized likelihood ratio test have been developed to reduce a possible loss of power due to an incorrect choice of the scanning window size. Monte Carlo simulations have been employed for calculating the critical value of all scan statistics discussed in this dissertation, as well as powers for selected parameters in the alternative hypotheses, indicating a local change in the scale parameter of the gamma distribution.

An application based on a real data example is also demonstrated for the use of multiple window and variable window scan statistics.

For a moderate to large shift of the scale parameter, in both one dimensional and two dimensional cases, simulation studies suggest that the conditional fixed window scan statistic performs well. When the true window size $m$ of the local change is unknown, both multiple window and variable window scan statistics perform well, demonstrating similar levels of test power. Regarding computational efficiency, however, the variable window scan statistic would generally take longer than the multiple window scan statistic to implement. Therefore, if the sample size of the total scanning region is relatively small, variable window scan statistic is recommended; for a larger sample size or a higher dimensional case, multiple window scan statistic is suggested with a faster implementation. However, one needs to be cautious of the choice of scanning windows used in the multiple window scan statistic, in order to obtain a wide coverage and a relatively close estimate of the true window size. Furthermore, both multiple window and variable window scan statistics can estimate the size and location of the local change, with variable window scan statistic reporting more accurate results. We also observe higher power for gamma random variables with a larger shape parameter, suggesting ore accurate detection of the local change. Compared to the one dimensional case, the test powers of all scan statistics discussed are relatively high at a moderate shift of the scale parameter.

## 5.2 Future Work

When the observations can be modeled by a gamma distribution $\Gamma(\theta, \beta)$, where $\theta$ is known to be an integer or estimated to be close to an integer value, they can be considered as the sum of multiple exponential random variables following $\text{Exp}(\beta)$. Therefore, many problems for gamma random variables could be potentially solved based on solutions for exponential distribution. Exploring different topics on scan statistics for exponential random random variables would be a possible future direction of research.

Another interesting topic is to change our alternative hypotheses from $\beta_1 > \beta_0$ to $\beta_1 - \beta_0 > \epsilon$, where both $\beta_0$ and $\epsilon$ are known. In other words, it is only of interest to detect a local change that exceeds a certain threshold in the scale parameter. This problem has many potential applications, especially in quality control.

While we have discussed scan statistics for a two dimension array, there is also interest in three dimensional scan statistics, which could be a prospective direction. Other future work related to this dissertation includes investigation on detecting a relatively small local shift of the scale parameter given the shape parameter is known, and the development in algorithms that can further improve the computational efficiency of multiple window and variable window scan statistics, especially when they are applied to larger scanning regions. Another direction for future research is to develop scan statistics for detecting a local change in the scale parameter, when the shape parameter is constant but unknown.

# Bibliography


Aksoy, H. (2000). Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences 24*(6), 419–428.

Alm, S. E. (1997). On the distributions of scan statistics of a two-dimensional poisson process. *Advances in Applied Probability 29*(1), 1–18.

Alm, S. E. (1998). Approximation and simulation of the distributions of scan statistics for poisson processes in higher dimensions. *Extremes 1*(1), 111–126.

Balakrishnan, N. and V. B. Nevzorov (2004). *A Primer on Statistical Distributions*. John Wiley & Sons, New York.

Boland, P. J. (2007). *Statistical and probabilistic methods in actuarial science*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, New York.

Chen, J. and J. Glaz (1996). Two-dimensional discrete scan statistics. *Statistics & Probability Letters 31*(1), 59–68.

Chen, J. and J. Glaz (2016). Scan statistics for monitoring data modeled by a negative binomial distribution. *Communications in Statistics–Theory and Methods 45*(6), 1632–1642.

Chen, J. and J. Glaz (2017). Scan statistics for integer-valued random variables: Conditional case. In J. Glaz and M. V. Koutras (Eds.), *Handbook of Scan Statistics*, pp. 1–32. Springer, New York.

Daley, D. and D. Vere-Jones (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer, New York.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*, Volume 1. Cambridge University Press, New York.

Devroye, L. (2013). *Non-Uniform Random Variate Generation*. SpringerLink : Bücher. Springer, New York.

Gan, F. (1998). Designs of one-and two-sided exponential ewma charts. *Journal of Quality Technology 30*(1), 55–69.

Glaz, J. and N. Balakrishnan (2012). *Scan Statistics and Applications*. Statistics for Industry and Technology. Birkhäuser, Boston.

Glaz, J. and J. Naus (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *The Annals of Applied Probability 1*(2), 306–318.

Glaz, J., J. Naus, and X. Wang (2012). Approximations and inequalities for moving sums. *Methodology and Computing in Applied Probability 14*(3), 597–616.

Glaz, J., V. Pozdnyakov, and S. Wallenstein (2009). *Scan Statistics: Methods and Applications*. Birkhäuser, Basel.

Glaz, J. and Z. Zhang (2004). Multiple window discrete scan statistics. *Journal of Applied Statistics 31*(8), 967–980.

Haiman, G. (2007). Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. *Journal of Statistical Planning and Inference 137*(3), 821–828.

Haiman, G. and C. Preda (2002). A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process. *Methodology and Computing in Applied Probability 4*(4), 393–407.

Haiman, G. and C. Preda (2006). Estimation for the distribution of two-dimensional discrete scan statistics. *Methodology and Computing in Applied Probability 8*(3), 373–382.

Hoh, J. and J. Ott (2000). Scan statistics to scan markers for susceptibility genes. *Proceedings of the National Academy of Sciences 97*(17), 9615–9617.

Jarrett, R. (1979). A note on the intervals between coal-mining disasters. *Biometrika 66*(1), 191–193.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics–Theory and Methods 26*(6), 1481–1496.

Levin, A. M., D. Ghosh, K. R. Cho, and S. L. Kardia (2005). A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics 21*(12), 2867–2874.

Maguire, B. A., E. Pearson, and A. Wynn (1952). The time intervals between industrial accidents. *Biometrika 39*(1/2), 168–180.

Mendoza-Parra, M.-A., M. Nowicka, W. Van Gool, and H. Gronemeyer (2013). Characterising chip-seq binding patterns by model-based peak shape deconvolution. *BMC genomics 14*(1), 834.

Nagawalla, N. (1996). A scan statistic with a variable window. *Statistics in Medicine 15*(7-9), 845–850.

Naus, J. (1965). Clustering of random points in two dimensions. *Biometrika 52*(1-2), 263–266.

Naus, J. (1966). Power comparison of two tests of non-random clustering. *Technometrics 8*(3), 493–517.

Wang, X. and J. Glaz (2014). Variable window scan statistics for normal data. *Communications in Statistics–Theory and Methods 43*(10-12), 2489–2504.

Zhang, C., M. Xie, J. Liu, and T. Goh (2007). A control chart for the gamma distribution as a model of time between events. *International Journal of Production Research 45*(23), 5649–5666.

Zhao, B. and J. Glaz (2016). Scan statistics for detecting a local change in variance for normal data with unknown population variance. *Statistics & Probability Letters 110*, 137–145.

Zhao, B. and J. Glaz (2017). Scan statistics for detecting a local change in variance for two-dimensional normal data. *Communications in Statistics–Theory and Methods 46*(11), 5517–5530.