5-20-2015

Scan Statistics for Detecting a Local Change in Variance for Normal Data

Bo Zhao
bo.zhao@uconn.edu

Scan Statistics for Detecting a Local Change in Variance for Normal Data

Bo Zhao, Ph.D.
University of Connecticut, 2015

In this dissertation scan statistics for detecting a local change in variance are proposed for both one and two dimensional normal observations. When the size of the window where a local change has occurred is known, fixed window scan statistics are proposed. Approximations for the distributions of fixed window scan statistics are investigated. When the correct window size is unknown, variable window scan statistics based on generalized likelihood ratio tests and multiple window minimum P-value scan statistics are developed. When population variance, where the null hypotheses of no change in variance, is also unknown, a conditional approach is employed, for the proposed method of implementing the scan statistics. Conditional variable and multiple window scan statistics are also derived in case both the variance and the window size are unknown. For moderate or large shift in variance, multiple and variable window scan statistics performed well.

Scan Statistics for Detecting a Local Change in Variance for Normal Data

Bo Zhao
B.S., Statistics, Nanjing University, China, 2009

A Dissertation

Submitted in Partial Fullfilment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Scan Statistics for Detecting a Local Change in Variance for Normal Data

Presented by
Bo Zhao, B.S.

Major Advisor _____
Joseph Glaz

Associate Advisor _____
Nitis Mukhopadhyay

Associate Advisor _____
Vladimir Pozdnyakov

University of Connecticut

2015

To my parents Genxi, Lan and my wife Lu

ACKNOWLEDGEMENTS

First, I would like to sincerely thank Professor Joseph Glaz, for all his guidance, patience and support to me in the past few years. And I always consider myself extremely lucky to have him not only being my advisor for academic research, but also as a great mentor and friend in life, who always remind me the warmth and beauty of the real world outside of the cold books.

Moreover, I would like to thank Professor Nitis Mukhopadhyay and Professor Vladimir Pozdnyakov, for being my associate advisors in my dissertation committee, and for all the great advice that enriched and improved my dissertation. I also would like to thank Professor Richard Vitale, Professor Zhiyi Chi, Professor Minghui Chen and all other professors, staff and friends here in the department of statistics, my journey here would be much tougher without all your helps and smiles. And I would like to thank the department of statistics for providing me four years of assistantship, and a great place to learn and do research. [1]

Last but not least, I would like to thank my parents Genxi, Lan and my wife Lu, for their love and support all these years in my life adventures.

[1] Part of the computation in this dissertation was done on the Beowulf cluster of the Department of Statistics, University of Connecticut, partially financed by the NSF SCREMS (Scientific Computing Research Environments for the Mathematical Sciences) grant number 0723557.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Introduction

Scan statistics for observations in a one or two dimensional region, have been of great interest and have risen to prominence in applied probability and statistics in the last 25 years. This is due to numerous applications in a wide varieties of fields, including: archaeology, astronomy, bio-informatics, bio-surveillance, computer science, electrical engineering, epidemiology, food sciences, genetics, geography, materials sciences, molecular biology, physics, reconnaissance, reliability and quality control, and telecommunication etc. ([18], [14], [15], [2], [11], [16], [20], [1], etc.)

Scan statistics are effectively used to detect a local change in a parameter of the distribution, and are based on a sequence of moving windows in one dimensional, or two dimensional region. When the correct window size within the monitored data, where a change in the parameter has occurred is known, a fixed scanning window can be employed for the scanning procedure. However, in practice this window size will often not be available. Hence, to avoid loss of power by using an incorrect window size for the local detection of the parame-

ter change, several scanning windows of different length need to be incorporated in our scanning procedure. The former case is naturally named the fixed window scan statistic in the literature, and the latter, will be referred to as variable window scan statistics or multiple window scan statistics.

Most of the research in the area of scan statistics has been focused on detecting a local change in the mean of the underlying process, employing both fixed window scan statistics ([15], [21], [22]), multiple window scan statistics ([24], [17] and [40]) and variable window scan statistics ([30], [32] and [29,28]). For normal data, [19] and [39] investigated the performance of fixed and multiple window scan statistics for detecting a local change in mean. Not much research has been carried out in the scientific literature for detecting a local change in variance, which may have potential applications in financial risk monitoring ([35], etc.), among other areas. A related problem for detecting a continual change in variance, employing methods from the area of sequential analysis, has been investigated by several researchers, including: [25], [26], [34], [38], [36], and [33].

This dissertation research is focusing on scan statistics for detecting a local change in variance for normal data. An interesting approach for detecting a change of variance by employing moving sum of squares statistics for normal data was proposed by Bauer and Hackle in [3,4], in which the focus is on sequential detection of a persistent shift of the mean and variance. Moreover, the dependence structure of moving sums of squares statistics is not utilized and their joint

distributions are approximated only by marginal distributions. Some other major references that can provide additional motivation for our problem include variable window scan statistics via a generalized likelihood ratio approach based on discrete distributions derived by Nagarwalla in [32] and Kulldorff in [29]; approximations for probabilities of moving-sum-type scan statistics in [19], [21], etc.; a conditional approach for scan statistics based on the negative binomial distribution when the baseline parameter is unknown in [6]; and a thorough discussion for the scan statistics for detecting a local change in mean level for normal data in [39], based on which I will extend the topic to detecting a local change in variance for normal data as my dissertation research.

The rest of the dissertation is organized as follows: Chapter 2 formally describes the hypotheses tests of interest in one dimensional case, and proposes the fixed, variable and multiple window scan statistics assuming the population variance under the null hypothesis is known. In Chapter 3, three approaches are proposed to perform our scan tests when the variance under the null hypothesis is unknown: a training sample approach, a conditional approach and a parametric bootstrap testing approach. A conditional variable window scan statistic is also derived and compared to a conditional multiple window scan statistic in Chapter 3, in case both the cluster size and the population variance is unknown. Our one-dimensional scan statistics are extended to two dimensional cases in Chapter 4, for both the cases of variance known and unknown in the null hypothesis. A

brief summary of results and proposal of future works for this dissertation is given in Chapter 5.

Chapter 2

Scan Statistics for One-dimensional Normal Data with Known Variance

2.1 Introduction

A major interest of scan statistics is detecting a local change in a parameter of the model for the observed data. In this chapter we investigate the use of scan statistics for detecting a local change in variance for sequence of observations.

Let $X_1, \ldots, X_M$ be a sequence of independent and identically distributed (i.i.d.) normal observations with mean μ and variance σ^2, where M is the specified range of the monitoring process. We are interested in detecting a local upward shift in variance. One can modify the methods in this dissertation easily to detect also a local downward shift or a two-sided shift. Let $2 \leq m \leq M/4$, be the size of the sliding window of a segment of m consecutive observations, we are interested in testing the following hypotheses: H_0: $X_i, 1 \leq i \leq M$, are i.i.d. normal random variables with mean μ and variance σ_0^2, vs. H_a: $X_i, 1 \leq i \leq M$, are independent normal random variables with mean μ, and the $X_i's$, have variance $\sigma_1^2 > \sigma_0^2$ for

$i \in R(a,m) = \{a, a+1, ..., a+m-1\}$, for some a such that $1 \leq a \leq M-m+1$. but the $X_i's$ for $i \notin R(a,m)$ have variance σ_0^2. This a is the unknown starting location of the window where a local change in variance has occurred. The restriction $m \leq M/4$ is used to emphasize that in most applications, we are interested in detecting a local change in variance within a window of small or moderate size with length no more than $M/4$, but when necessary, one could certainly consider m for larger possible values, say $M/2$, etc.

In the above hypotheses one can always assume that $\mu = 0$. If $\mu \neq 0$, one can replace the $X_i's$ with the sequence of recurrent residuals:

$$W_i = \frac{(i-1)X_i - \sum_{j=1}^{i-1} X_j}{\sqrt{i(i-1)}}, 2 \leq i \leq M \tag{2.1}$$

which are i.i.d normal random variables with mean 0 and variance σ_0^2, under the null hypothesis. ([3]). In fact, W_i corresponds to the ith entry of the so-called *Helmert* transformation of the original data sequence ([31,p.197]). The use of this transformation will result in losing one observation.

When σ_0^2 is known, without loss of generality one can assume $\sigma_0^2 = 1$. Therefore, the testing problem reduces to testing H_0: $X_i, 1 \leq i \leq M$, are i.i.d. normal random variables with mean 0 and variance 1 vs. H_a: $X_i, 1 \leq i \leq M$, are independent normal random variables with mean 0, the $X_i's$ have variance $\sigma_1^2 > 1$ for $i \in R(a,m)$ and $\sigma_0^2 = 1$ for $i \notin R(a,m)$. In this chapter, we only focus on the case when σ_0 is assumed known, and for simplicity, we assume $\sigma_0^2 = 1$.

The rest of this chapter discusses our scan statistics and related issues for

our problem formalized above, and is organized as follows. In Section 2.2 and 2.3, we discuss the fixed window scan statistic and two approximations for the distribution of the fixed window scan statistic respectively. For a given significance level, we present a searching algorithm in Section 2.4 to evaluate the critical value for our testing problem. In Section 2.5, we derive a variable window scan statistic via the generalized likelihood ratio method. We present an algorithm for the implementation of this test statistic. In Section 2.6, a multiple window minimum p-value scan statistic is developed along with an algorithm for its implementation. In Section 2.7, for selected values of the parameters, we present numerical results to evaluate the performance of the scan statistics investigated in this chapter. Concluding remarks are given in Section 2.8.

2.2 A Fixed Window Scan Statistic

Let $2 \leq m \leq M/4$ be a prespecified length of the sliding window. A fixed window *scan statistic* for detecting a local change in variance, is defined by:

$$S_{m,M} = max\{Y_{r,m}; 1 \leq r \leq M - m + 1\}, \tag{2.2}$$

where $Y_{r,m}$ are the moving sums of squares of the observed data:

$$Y_{r,m} = \sum_{i=r}^{r+m-1} X_i^2; 1 \leq r \leq M - m + 1. \tag{2.3}$$

Under H_0, the random variables $Y_{r,m}, 1 \leq r \leq M - m + 1$, are m-dependent and have a joint multivariate chi-square distribution and marginal chi-square dis-

tributions with m degrees of freedom. The joint covariance matrix is given by: $\Sigma = \{\sigma_{i,j}\}$, where: $\sigma_{i,i} = 2m$, for $1 \leq i \leq m$, $\sigma_{i,j} = 0$, for $|j - i| \geq m$ and $\sigma_{i,j} = 2(m - k)$, for $|j - i| = k$, $1 \leq k \leq m - 1$.

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

$$G_{m,t}(M) = P(S_{m,M} < t) = P(Y_{1,m} < t, Y_{2,m} < t, \ldots, Y_{M-m+1,m} < t), \qquad (2.4)$$

be the cumulative distribution function of $S_{m,M}$. Then,

$$P(S_{m,M} \geq t) = 1 - G_{m,t}(M). \qquad (2.5)$$

When the values of m, M and t are clearly understood, we abbreviate $G_{m,t}(M)$ and $S_{m,M}$ to $G(M)$, and S_m, respectively. For our hypotheses testing problem, when the window size m is known, the generalized likelihood ratio test rejects the null hypothesis, in favor of the local change alternative hypothesis H_a, whenever $S_{m,M}$ exceeds a threshold value t, where t is determined by $P(S_{m,M} \geq t|H_0) = \alpha$, α being the specified significance level. Hence, to implement our testing procedure we need to evaluate $G(M)$.

Unlike the case of detecting a local change in the mean level for the normal data, where extensive theoretical results and R algorithms for computing multivariate normal and t distributions are readily available ([13] and [39]), for the problem at hand there are no algorithms to evaluate $G(M)$. Numerous types of multivariate chi-square and gamma distributions are discussed in [27], and other references as well. None of these results are applicable to our data structure.

Due to the complexity of the dependence structure of the multivariate chi-square distribution for $Y_{r,m}, 1 \leq r \leq M - m + 1$, one has to rely on direct Monte Carlo simulation for computing approximations for $G(M)$.

2.3 Approximations for Probabilities of the Fixed Window Scan Statistic

We now present two approximations for $G(M)$. It follows from [19], Equation (14), that:

$$G(M) = G(3m) \left[\frac{G(3m)}{G(2m)}\right]^{K-3} \frac{G(2m+v)}{G(2m)}, \tag{2.6}$$

where $K \geq 3, m \geq 2$ and $0 \leq v \leq m-1$ are integers such that $M = Km + v$. The second approximation for $G(M)$ is based on [21], Corollary 2 and Equation (2.2):

$$G(M) = \frac{2G(2m) - G(3m)}{[1 + G(2m) - G(3m) + 2(G(2m) - G(3m))^2]^{M/m-1}}, \tag{2.7}$$

where a sharp approximation of the error bound is given by:

$$3.3[1 - G(2m)]^2(M/m - 1), \tag{2.8}$$

provided $M \geq 3m$, $1 - G(2m) \leq 0.025$ and $3.3M[1 - G(2m)]^2(M/m - 1) \leq 1$. The two approximations reduce significantly the computing time needed to evaluate the multivariate chi square distributions, especially when M/m is large. We employ direct Monte Carlo simulation for computing $G(3m)$, $G(2m)$, and $G(2m + v)$, to evaluate the two approximations given above. Notice that, both approximations are also valid for i.i.d. observations from distributions other than

normal. In Section 2.7, for selected values of the parameters, we evaluate the accuracy of these approximations.

2.4 An Algorithm for Rejection Threshold Searching

To implement the fixed window scan statistic, $S_{m,M}$, for testing the hypotheses outlined above, one has to determine the rejection region for a specified significance level α. We now present an effective algorithm for searching for the critical value that determines the rejection region. This algorithm is based on the approximations in Equation (2.6) or (2.7).

For given values of M, m, α and a searching precision parameter r the steps of this algorithm are:

- Find integers K and v, such that $M = Km + v$.
- Run N simulations, each generating $3m$ i.i.d. $N(0,1)$ random variables.
- For the ith simulation ($i = 1, \ldots, N$), compute the value of the scan statistic by definition (2.2), based on sequences starting from the beginning of the simulation and of length $M = 2m, 3m, 2m+v$, denoted by $s_{i,2m}, s_{i,3m}, s_{i,2m+v}$, respectively.
- To search for the critical value t, such that $P(S_{m,M} > t) = \alpha$, start searching in the region $[L, U]$. Usually the region $[0, 4m]$ will perform well. A maximum searching time T can be set *a priori*.

1. Compute $t = (L + U)/2$.
2. For $d = 2m, 3m, 2m+v$, compute the percentages of $\{s_{i,d}, i = 1, \ldots, N\}$ that is smaller than t and denote it as $G(d)$.
3. Compute type I error probability P by (2.5) via the approximation in Equation (2.6) or (2.7).
4. If $|P-\alpha| < r$, stop and return t as the critical value; otherwise proceed as follows:
   - If $P > \alpha$ let $L = t$, $U = U$ and restart from step 1.
   - If $P < \alpha$ let $L = L$, $U = t$ and restart from step 1.
5. If maximum searching time T is reached and the precision is not obtained, give a warning.

In Section 2.7, for selected values of the parameters, we evaluate the performance of this algorithm via simulation.

2.5 A Variable Window Scan Statistic via the Generalized Likelihood Ratio Method

One limitation in using a fixed window scan statistic is that most often in practice one does not know the exact length m of a window where a local change of the variance has occurred. When the length of the scanning window is far from the actual length of the sequence of data, where a change in the variance has

occurred, the power of a fixed window scan statistic will be greatly reduced, One approach to solve this problem is to derive a variable window scan statistic via the generalized likelihood ratio test (GLRT) method, following the approach in [32] and [29], where detection of a local change in the mean of observed data has been investigated.

We now outline the steps for deriving this variable window scan statistic. For a local upward shift in variance, the generalized likelihood ratio test will reject H_0 in favor of H_a for large values of

$$\Lambda = \frac{\sup_{\theta \epsilon \Theta_1} \prod_{i=1}^{M} f_\theta(x_i)}{\sup_{\theta \epsilon \Theta_0} \prod_{i=1}^{M} f_\theta(x_i)}, \tag{2.9}$$

where $f_\theta(x_i)$ is the probability density of the ith observation in the scanned sequence $\{X_i\}$ and Θ_0 and Θ_1 are the parameter spaces for the null and alternative hypotheses, respectively. This generalized likelihood ratio statistic can be expressed explicitly as:

$$\begin{aligned}
\Lambda &= \sup_{\Theta_1} \left(\frac{1}{\sigma_1}\right)^m exp\left(\frac{1}{2}\sum_{i=a}^{a+m-1} X_i^2 - \frac{1}{2\sigma_1^2}\sum_{i=a}^{a+m-1} X_i^2\right) \\
&= \sup_{\Theta_1} \left(\frac{1}{\sigma_1}\right)^m exp\left(\frac{1}{2}Y_{a,m} - \frac{1}{2\sigma_1^2}Y_{a,m}\right) \\
&= \sup_{a;m} \left(\frac{m}{Y_{a,m}}\right)^{m/2} exp\left(\frac{1}{2}Y_{a,m} - \frac{m}{2}\right),
\end{aligned} \tag{2.10}$$

where $Y_{a,m} = \sum_{i=a}^{a+m-1} X_i^2$. The last step follows from the fact that for fixed but arbitrary a and m, constrained by parameter space Θ_1, the supremum is achieved

at $\hat{\sigma}_1^2 = Y_{a,m}/m > 1$. Let

$$L_m(Y_{a,m}) = \left(\frac{m}{Y_{a,m}}\right)^{m/2} exp\left(\frac{1}{2}Y_{a,m} - \frac{m}{2}\right). \tag{2.11}$$

Regard $L_m(Y_{a,m})$ as a function of $Y = Y_{a,m}$, depending on a, for fixed but arbitrary m. This function is a convex function of Y and it is increasing in Y on Θ_1. Therefore, for fixed m, the supremum in (2.10) is achieved at the maximum value of Y. One can obtain a unique value of a that maximizes Y. It follows that, for a given sequence of observations, one can get the location and length of the window that maximizes $L_m(Y_{a,m})$. This maximum value of $L_m(Y_{a,m})$ is the value of our variable window scan statistic based on the generalized likelihood ratio principle. For a given sequence of observations $X_1, \ldots, X_M$ in our testing problem, the algorithm presented below implements the search for the location and length of the window that maximizes $L_m(Y_{a,m})$.

- For $2 \leq m \leq M/4$, execute the following steps.
  - Compute $Y_{a,m} = \sum_{i=a}^{a+m-1} X_i^2$ for all a, where $1 \leq a \leq M - m + 1$.
  - Find $\max\{Y_{a,m}; 1 \leq a \leq M - m + 1\}$ and record as $Y^*(m)$ and the corresponding a record as $a^*(m)$.
  - Compute $L^*(m) = L_m(Y^*(m))$ by equation (2.11).
- Find $\max_{2 \leq m \leq M/4}\{L^*(m)\}$, record it as Λ^* and record the corresponding m as m^*. Then Λ^* is the value of the variable window scan statistic; m^* is the

most likely window size where a possible upward local change in variance has occurred and $a^*(m^*)$ is the most likely starting position for the local change.

The p-value corresponding to the observed value of Λ^* can be found by a simulation algorithm as follows:

- Perform N simulations, each generating M i.i.d. $N(0,1)$ random variables.
- For each simulation, compute the observed value of Λ in the same way as Λ^* has been evaluated by the algorithm presented above.
- The P-value is equal to the proportion of the observed values of Λ's, based on the N simulations, that exceeds Λ^* for the data set

In Section 2.7, for selected values of the parameters, we evaluate the performance of the variable window scan statistic based on the generalized likelihood ratio principle.

2.6 A Multiple Window Scan Statistic via the Minimum P-value Approach

In the previous section, a variable window scan statistic based on the generalized likelihood ratio principle has been discussed. It involved scanning the data with windows of length m, where $2 \leq m \leq M/4$. If M is large this procedure might

become computationally intensive. For the problem at hand, we propose to investigate the performance of the following multiple window scan statistic, based on the minimum P-value method ([17] and [39]).

Since the window length m, where the change in the variance has occurred is unknown, a sequence of n fixed window scan statistics $\{S_{m_1}, S_{m_2}, \ldots, S_{m_n}\}$ can be employed simultaneously, where $2 \leq m_1 < m_2 < \ldots < m_n \leq M/4$. The lengths of the n sliding windows are chosen in advance by the experimenter. For $1 \leq j \leq n$, let t_j be the observed value of S_{m_j} and $p_j = P(S_{m_j} > t_j | H_0)$ its associated p-value. To test H_0 vs. H_a, the minimum p-value statistic, P_{min}, is defined as follows:

$$P_{min} = \min\{p_j; 1 \leq j \leq n\}. \tag{2.12}$$

The null hypothesis is rejected if the observed value of P_{min} falls below a critical value corresponding to a specified significance level α. Since the exact distribution of the P_{min} statistic is unknown, for a given significant level α, the critical value p_α,

$$P_{H_0}(P_{min} < p_\alpha) = \alpha, \tag{2.13}$$

has to be evaluated by a Monte Carlo simulation. The following algorithm can be used to find the critical value p_α:

- Run N simulations with M i.i.d. $N(0,1)$ observations generated in each run of the simulation.

- For each run of the simulation, record the observed values of the fixed window scan statistics, $S_{m_1}, \ldots, S_{m_n}$, denoted by $t_1, \ldots, t_n$, respectively. For $1 \leq j \leq n$, compute the observed p-value, $p_j = P_{H_0}(S_{m_j} > t_j)$, from the simulations obtained in previous steps.

- For each run of the simulation, record the minimum value of $p_j, 1 \leq j \leq n$, as p_{min}.

- p_α will be the $(\alpha \times 100)$th percentile of the N p_{min}'s.

In Section 2.7, for selected values of the parameters, we evaluate the performance of the P_{min} statistic. We present a simulation study to compare the power of variable, multiple and fixed window scan statistics.

2.7 Numerical Results

We first present numerical results to evaluate the accuracy of the approximations for $P(S_m > t)$, in Equations (2.6) and (2.7), based on simulated probabilities of $G(2m)$ and $G(3m)$ from $100,000$ trials. In Tables 2.1 and 2.2, for selected values of M, m and t, these two approximations are compared with probabilities for $P(S_m > t)$, computed by a direct simulation with $100,000$ trials of i.i.d. observations of $N(0,1)$. From the numerical results, it is evident that the two approximations are quite accurate.

Table 2.1: Approximations of $P(S_m > t)$ for 1-dim Scan Statistics, $M = 200$, $m = 5$

t	20	21	22	23	24	25
Direct Simulation	0.112110	0.075210	0.050960	0.033720	0.022170	0.014610
Glaz's Approximation	0.106905	0.072632	0.048619	0.033395	0.019635	0.012233
Haiman's Approximation	0.106744	0.072564	0.048593	0.033383	0.019629	0.012230

Table 2.2: Approximations of $P(S_m > t)$ for 1-dim Scan Statistics, $M = 500$, $m = 10$

t	32	33	34	35	36	37
Direct Simulation	0.066810	0.046610	0.032570	0.022240	0.015100	0.010550
Glaz's Approximation	0.071175	0.047460	0.031893	0.019521	0.016510	0.010190
Haiman's Approximation	0.071166	0.047459	0.031892	0.019519	0.016511	0.010191

Table 2.3: Rejection Thresholds for 1-dim Scan Statistics by Searching Algorithm with Approximation or Direct Simulation

Significance Level α	Searching Method	Rejection Threshold m=5	m=10	m=15	m=20	m=25
0.01	A	24.61	33.05	40.61	47.46	54.49
	S	24.22	32.91	40.50	47.45	54.20
0.05	A	20.21	28.30	35.07	41.66	48.12
	S	20.28	28.18	35.11	41.63	47.90
0.10	A	18.37	26.06	32.56	38.94	44.93
	S	18.53	26.00	32.65	38.93	44.89

For selected significance levels α, Table 2.3 provides numerical results for the rejection thresholds of fixed window scan statistics, with scanning window lengths $m = 5, 10, 15, 20, 25$, respectively, for a sequence of $M = 100$ $N(0, 1)$ observations. The thresholds in each entry of Table 2.3 are obtained either by direct simulations, marked as method "S", or by the searching algorithm, marked as method "A", based on the approximation in Equation (2.6). The direct simulation method is based on $N = 100,000$ trials. The searching algorithm is based on the parameters $r = 0.001$ and $T = 10,000$. Based on the numerical results, the rejection thresholds obtained by the two methods in Table 2.3 are very close, indicating good performance of our searching algorithm.

Table 2.4: Power Comparison for 1-dim Fixed, Multiple and Variable Window Scan Statistics, σ_0^2 Known, $M = 100$

	$\sigma_1/\sigma_0 = \sqrt{2}$			$\sigma_1/\sigma_0 = 2$			$\sigma_1/\sigma_0 = 3$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
5	0.1121	0.2349	0.3319	0.6031	0.7368	0.8014	0.9561	0.9772	0.9850
10	0.1258	0.2608	0.3604	0.6499	0.7761	0.8316	0.9676	0.9835	0.9889
15	0.1174	0.2538	0.3535	0.6236	0.7587	0.8207	0.9621	0.9807	0.9876
20	0.1066	0.2394	0.3398	0.5940	0.7367	0.8033	0.9559	0.9775	0.9851
25	0.0943	0.2181	0.3208	0.5569	0.7054	0.7798	0.9471	0.9724	0.9817
MW	0.1195	0.2545	0.3523	0.6342	0.7648	0.8219	0.9645	0.9816	0.9879
VW	0.1219	0.2555	0.3497	0.6394	0.7662	0.8218	0.9652	0.9818	0.9880

Table 2.5: Power Comparison for 1-dim Fixed, Multiple and Variable Window Scan Statistics, σ_0^2 Known, $M = 250$

	$\sigma_1/\sigma_0 = 2$			$\sigma_1/\sigma_0 = \sqrt{7}$			$\sigma_1/\sigma_0 = 3$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
5	0.5364	0.6707	0.7389	0.8736	0.9263	0.9459	0.9404	0.9665	0.9766
15	0.5612	0.6949	0.7579	0.8892	0.9344	0.9520	0.9503	0.9724	0.9799
25	0.5085	0.6541	0.7284	0.8625	0.9190	0.9409	0.9356	0.9646	0.9748
MW	0.5671	0.6952	0.7608	0.8910	0.9348	0.9531	0.9510	0.9719	0.9806
VW	0.5759	0.7036	0.7660	0.8955	0.9379	0.9547	0.9528	0.9735	0.9813

In Tables 2.4 and 2.5, for selected values of the parameters and three different scales of the local shift in variance, represented by the ratio of σ_1/σ_0, we present numerical results for power of fixed, variable and multiple window scan statistics, based on a simulation with $N = 100,000$ trials of $N(0,1)$ observations. In Table 2.4 a sequence of length $M = 100$ observations is used, with a local change of variance starting at the 11th observation, for a length of $m = 10$ consecutive observations. In Table 2.5, we used a sequence of length $M = 250$, with a local shift in variance starting at 101th observation and $m = 10$. Since we assume a potential local change in variance, the variable window (VW) scan statistics, in the two tables employed $3 \leq m \leq 25$ and $3 \leq m \leq 50$, respectively. The multiple window (MW) scan statistics used all the listed fixed window sizes simultaneously. Note that the true window size for a local change in Table 2.5 is not among the selected window sizes for fixed or multiple window scan statistics.

From the numerical results in Tables 2.4 and 2.5, for each of the shift ratios σ_1/σ_0, it is evident that the power is maximized by using a fixed window scan statistic with the correct window size. When the correct length of the window where a change in the variance has occurred is unknown, using a fixed window scan statistic with an incorrect window size most often will result in sizable loss of power. The power loss can be greatly reduced by employing a variable or a multiple window scan statistic. Based on our power simulation study, the variable window scan statistic outperformed slightly the multiple window scan statistic.

Simulations for sequences of length $M = 1000$ yielded similar results.

2.8 Concluding Remarks

We have investigated the performance of several types of scan statistics for detecting a local change in variance for one dimensional i.i.d. normal observations with known variance. When the local shift in variance is moderate or large and the size of the window, where the change has occurred, is unknown, the variable and multiple scan statistics performed well. The variable window scan statistic performed slightly better than the multiple window scan statistic. When the length M of the scanned data sequence gets large, the implementation of the multiple window scan statistic is much faster than that of the variable window scan statistic, with only a small loss of of power. Hence, for detecting a local change in variance in large data sets, the multiple window scan statistic is recommended.

We would like to mention that the variable window scan statistic can be potentially employed for detecting the most likely location and size of a local change of variance. Based on a simulation study with $100,000$ trials for a sequence of length $M = 100$, having a shift in standard deviation from 1 to 3, starting at 11th observation with window size $m = 10$, the percentage of the most likely location for a change, has been correctly detected within 2 units from the true starting point, is nearly 81.4%. The percentage of the most likely window size for the change, within 3 units from the true size, is nearly 75.1%. A similar simulation study with $M = 200$ and true window size $m = 30$, with the same parameters as in the previous example, showed that, the percentage of the most likely starting

point within 3 units from the true starting point, is nearly 88.3%. The percentage of the most likely length of the window where a change has occurred, within 5 units from the true size, is nearly 88.0%. Additional research needs to be done to investigate the performance of scan statistics for detecting accurately the location of a local change in variance and its length, when the change is moderate or small, and the sequence length is large.

Chapter 3

Scan Statistics for One-dimensional Normal Data with Unknown Variance

3.1 Introduction

In this chapter, we formulate and study the performance of fixed, multiple and variable window scan statistics for detecting a local change in variance for normal data when the population variance of the underlying normal distribution is unknown. In Chapter 2, scan statistics for detecting a local change in variance for one dimensional normal data have been discussed, with the assumption that the variance under the null hypothesis is known. However, in many practices the null variance is unknown, then the scan statistics proposed in Chapter 2 can not be applied directly, since the related computations rely on the known variance. In this chapter, three approaches are proposed to adjust the implementation of the tests when the null variance is unknown: a training sample approach, a conditional approach, and a parametric bootstrap testing approach.

The chapter is organized as follows. In Section 3.2, we formulate three fixed

window scan statistics for detecting a local change in variance for normal data. We present algorithms for approximating, via simulation, their distributions under the null hypothesis of no local change in variance and computing the power under specified alternatives. In Section 3.3, we develop a multiple window scan statistic, that can be viewed as a bootstrap test statistic, based on a sequence of observed P-values of a fixed window scan statistic discussed in Section 3.2. We also formulate a variable window scan statistic via a conditional generalized likelihood ratio test approach in Section 3.4. We present algorithms to evaluate the null distributions of these scan statistics and evaluate their power under specified alternatives. In Section 3.5, for selected values of the parameters, we present numerical results to evaluate the performance of scan statistics discussed in Sections 3.2 - 3.4. Concluding remarks are presented in Section 3.6.

3.2 Fixed Window Scan Statistics

3.2.1 A Training Sample Approach

Let $X_1, \ldots, X_M$ be a sequence of i.i.d. normal observations with mean μ and variance σ^2, where μ and σ^2 are unknown parameters and M is the specified range of the monitoring process. We are interested in detecting a potential occurrence of a local change in variance within a subsequence of m consecutive observations in the observed data. In this dissertation, we investigate the detection of a local upward shift in variance. One can modify the methods discussed in this dissertation to

accommodate detection of a local downward shift or a two-sided shift. Without loss of generality, one can always assume that $\mu = 0$. If $\mu \neq 0$, one can replace the $X_i's$ with the sequence of recurrent residuals:

$$W_i = \frac{(i-1)X_i - \sum_{j=1}^{i-1} X_j}{\sqrt{i(i-1)}}, 2 \leq i \leq M,$$

which are iid normal random variables with mean 0 and variance σ^2 ([3]). We are interested in testing H_0: $X_i, 1 \leq i \leq M$, are i.i.d. normal random variables with mean 0 and variance $\sigma^2 = \sigma_0^2$, vs. H_1: $X_i, 1 \leq i \leq M$, are independent normal random variables with mean $\mu = 0$, and variance $\sigma^2 = \sigma_1^2$ within a segment of m consecutive observations $R(a, m) = \{a, a+1, ..., a+m-1\}$, and variance $\sigma^2 = \sigma_0^2$ elsewhere, where $\sigma_1 > \sigma_0$. The parameter $1 \leq a \leq M - m + 1$, denotes the unknown starting location of the change in variance and $2 \leq m \leq M/4$ is the length of the window where a change in the variance has occurred. The restriction $m \leq M/4$ emphasizes the focus on detecting a local change in variance.

When m is known, one can employ a fixed window scan statistic:

$$S_{m,M} = max\{Y_{r,m}; 1 \leq r \leq M - m + 1\}, \tag{3.1}$$

where $Y_{r,m}$, is the moving sum of squares defined by:

$$Y_{r,m} = \sum_{i=r}^{r+m-1} X_i^2; 1 \leq r \leq M - m + 1. \tag{3.2}$$

One can show that the generalized likelihood ratio test rejects the null hypothesis in favor of the local change alternative whenever $S_{m,M}$ exceeds a threshold value t,

where t is determined by $P(S_{m,M} \geq t|H_0) = \alpha$, where α is a specified significance level of the testing procedure.

When σ_0^2 is known, Chapter 2 investigated the performance of this scan statistic for detecting a local change in variance. In most real applications the population variance is unknown. In such cases the methods in Chapter 2 cannot be applied to evaluate the tail probabilities $P(S_{m,M} \geq t|H_0)$. When σ_0^2 is unknown, the following three methods can be employed.

The first method is based on the availability of a training sample of n_0 i.i.d. observations, with an identical distribution as X_1 under H_0, and independent of the data $\{X_i, 1 \leq i \leq M\}$. Let $S_{n_0}^2$ be the sample variance evaluated from the training data. Then, under H_0,

$$\frac{(n_0 - 1)S_{n_0}^2}{\sigma_0^2}$$

has a $\chi_{n_0-1}^2$ distribution. Consider the transformed data:

$$X_i^* = X_i/S_{n_0}, 1 \leq i \leq M.$$

Under H_0, $X_i^*, 1 \leq i \leq M$, have a marginal t_{n_0-1} distribution and they are conditionally independent given $S_{n_0}^2 = s_{n_0}^2$. Note that when H_0 holds, the joint distribution of the new data sequence does not depend on any unknown parameters. We can employ the following scan statistic:

$$S_{m,M}^* = max\{Y_{r,m}^*; 1 \leq r \leq M - m + 1\}, \tag{3.3}$$

where

$$Y^*_{r,m} = \sum_{i=r}^{r+m-1} X_i^{*2} = \frac{\sum_{i=r}^{r+m-1} X_i^2}{S^2_{n_0}}; 1 \le r \le M - m + 1, \tag{3.4}$$

to test H_0 vs. H_1, stated above. The distribution function of $S^*_{m,M}$ is given by

$$G^*_{m,t}(M) = P(S^*_{m,M} < t) = P(Y^*_{1,m} < t, Y^*_{2,m} < t, \ldots, Y^*_{M-m+1,m} < t).$$

Under H_0, each $Y^*_{r,m}/m$ has an $F(m, n_0 - 1)$ distribution and $G^*_{m,t}(M)$ is a special type of a multivariate F distribution. Since there are no established algorithms to evaluate $G^*_{m,t}(M)$, the tail probabilities $P(S^*_{m,M} \ge t|H_0)$ have to be evaluated via a Monte Carlo simulation. Note that as $n_0 \to \infty$, the scan statistic probabilities $P(S^*_{m,M} \ge t|H_0) \to P(S_{m,M} \ge t|H_0)$. Therefore, when a training sample is available one can implement the testing of our hypotheses using a scan statistic for the transformed data $\{X^*_i, 1 \le i \le M\}$. In Section 3.5, Tables 3.1 and 3.2, we evaluate the performance of a fixed window scan statistic via the training sample approach.

3.2.2 A Conditional Approach

In some applications it may not be possible to obtain preliminary training data to estimate the population variance when H_0 is true. A second approach, that is widely applicable, to eliminate unknown parameters when H_0 is true, which for the problem at hand of the population variance, is to condition on the sufficient statistic under H_0. For example, this conditional approach is used for scanning negative binomial data in [6], where the joint distribution of the sequence of the

negative binomial data conditioning on their total sum was shown to have a multivariate Polya distribution, which does not depend on any unknown parameter. For our problem at hand, when H_0 is true, the sufficient statistic for σ_0^2 under H_0 is:

$$R^2 = \sum_{i=1}^{M} X_i^2.$$

In that case, the distribution of the random vector $\{X_i, 1 \leq i \leq M\}$, given $R^2 = r^2$, is uniform on a sphere of radius r. ([9,Chap.12], [37]). Moreover, the random vector

$$\{X_i^{**} = X_i/R; 1 \leq i \leq M\}, \tag{3.5}$$

where $R = \sqrt{\sum_{i=1}^{M} X_i^2}$, has a joint uniform distribution on the $(M-1)$ dimensional unit sphere. Consequently, we can define a scan statistic for the the sequence of observations $\{X_i^{**}; 1 \leq i \leq M\}$:

$$S_{m,M}^{**} = max\{Y_{r,m}^{**}; 1 \leq r \leq M - m + 1\}, \tag{3.6}$$

where

$$Y_{r,m}^{**} = \sum_{i=r}^{r+m-1} X_i^{**2} = \frac{\sum_{i=r}^{r+m-1} X_i^2}{R^2}; 1 \leq r \leq M - m + 1. \tag{3.7}$$

We propose to employ this scan statistic for testing H_0, conditional on $R^2 = r^2$. The conditional P-value of this scan statistic is given by

$$P(S_{m,M}^{**} \geq s | R^2 = r^2, H_0),$$

where s is the observed value of $S_{m,M}^{**}$. Under H_0, the distribution of $S_{m,M}^{**}$ does not depend on any unknown parameters. Hence, for a given significance level α

we can find the critical value t such that $P(S^{**}_{m,M} \geq t | R^2 = r^2, H_0) = \alpha$. These computations will be implemented via a Monte Carlo simulation that generates N sequences of data of M i.i.d. $N(0,1)$ observations, and then dividing each observation by R. In Section 3.5, Tables 3.1 and 3.2, we evaluate the performance of the fixed window scan statistic via the approach of conditioning on the sufficient statistic.

Alternative ways to make the data invariant of σ_0 similarly, include considering:

$$C_i = X_i/|X_1|, i = 2, \ldots, M$$

Then under the null hypothesis the C_i's marginally follows the standard Cauchy distribution, depending on each other but invariant to σ_0. Thus our scan statistics can also be implemented based on the $\{C_i\}$. Notice that, by using this transformation, a price to pay is losing one observation (only have $(M-1)$ C_i's), but this does not affect much for our test based on a long sequence.

3.2.3 A Parametric Bootstrap Testing Approach

A third approach for our testing problem is a parametric bootstrap test. As a popular computing intensive re-sampling method in statistics, the bootstrap idea was initially invented mainly for nonparametric estimation problems, e.g. estimating the confidence interval for the standard deviation of a test statistic $T(X)$ by re-sampling the original data $\{X_i; 1 \leq i \leq n\}$ with replacement many

times, to get the samples of the test statistics: $\{T_i\}$, etc.. But it has been adjusted quickly later for more widely applications, including hypothesis testing, regression problems, parametric settings, etc.

As pointed out by Efron in [10], the bootstrap method as a special data-based simulation relieves the analyst from having to make parametric distribution assumptions in nonparametric mode, and provides answers to non-textbook-formulae problems in parametric mode. For more details of the bootstrap methods. Detailed early development, theoretical justification and variations, etc, can be found in [10] and [8], etc.

For our problem at hand, the first step in implementing it is to estimate σ_0^2, the unknown population variance under H_0, via the sample variance of the observed data: $\widehat{\sigma_0^2} = S_M^2$. Let $\widehat{F}_0$ denote the fitted null model based on this estimate of σ_0^2. Then the calculation of the P-value for our fixed window scan statistic $S_{m,M}$, where s is its observed value,

$$p = P(S_{m,M} \geq s|\widehat{F}_0),$$

is referred to as a parametric bootstrap test. The P-value for this statistics has to be evaluated via simulation. The following algorithm can be effectively used to implement this parametric bootstrap test:

1. Given the sequence of observations $\{X_i; 1 \leq i \leq M\}$, and a scanning window of size m, compute the observed value of a fixed window scan statistic $S_{m,M}$

defined in (2.2), and denote it by $s_{m,M}$.

2. Based on $\{X_i; 1 \leq i \leq M\}$, estimate σ_0^2 by the sample variance S_M^2.

3. Generate B bootstrap samples indexed by b, $1 \leq b \leq B$, with each bootstrap sample consisting of a sequence of M i.i.d. normal observations with mean 0 and variance $\hat{\sigma_0^2} = S_M^2$.

4. For each generated bootstrap sample compute the fixed window scan statistic denoted by $S_{m,M}^{(b)}$, $1 \leq b \leq B$, defined in (2.2).

5. For a given significance level α, based on $S_{m,M}^{(b)}$, $1 \leq b \leq B$, compute the bootstrap P-value as follows:

$$\textit{P-value} = \frac{\#\{S_{m,M}^{(b)} > s_{m,M}; 1 \leq b \leq B\}}{B}. \tag{3.8}$$

   Reject the null hypothesis if the $\textit{P-value} < \alpha$.

To improve the precision of the bootstrap P-value obtained via this algorithm, one can use a double bootstrap procedure introduced by Davison and Hinkley in [8], which is computationally more intense. We present below the steps of an algorithm for the double bootstrap test.

1. Compute the P-value, p, based on the observed data and bootstrap samples as described in the algorithm above.

2. Repeatedly compute P-values, denoted by $p^{(b)}$, $1 \leq b \leq B$, for each of the bootstrap sample treated as original data and the re-sampled B bootstrap samples.

3. Then,

$$p_{adj} = \frac{\#\{p^{(b)} < p; 1 \leq b \leq B)}{B}$$

is the adjusted P-value.

To implement the parametric bootstrap test we need to generate B bootstrap samples from the estimated distribution under null hypothesis. A natural question is: how many of such bootstrap samples will be enough for our test to have an accurate estimate for the P-value and good power? An answer to this question can be found in [7], among many others, where a general rule for choosing the bootstrap sample size B is presented to keep the power loss at an acceptable level. It is suggested that B should be at least 400, preferably 1000, for significance level 0.05. It should be be at least 1500, preferably 4000, for level 0.01. In this article we have chosen $B = 1500$.

When necessary, one can also use a pre-testing algorithm given in [7] for searching B, which showed very good performance supported by simulation experiments. The brief idea for this pre-testing algorithm is described as follows: for a given significance level α, start with a small value of B, then,

1. Do regular bootstrap test and compute bootstrap P-value as in (3.8);

2. Utilize a separate test as follows to check if the bootstrap sample size B is large enough: If the obtained P-value in step 1 is significantly smaller (or larger) than the specified significance level α, then stop and conclude that B is good enough, else add more bootstrap samples and repeat the procedure until stop or a maximum-allowed value of B is reached.

   The above significance testing for P-value smaller (or larger) than α is done by testing $H_1 : P\text{-}value < \alpha$ (or $H_1 : P\text{-}value > \alpha$), using binomial distribution or its normal approximation with another pre-specified significance level β. Refer to [7] for more details.

To compare the performance of the three scan statistics discussed in this Section, we need to evaluate and compare their power for specified local alternatives. For a specified significance level α and a specified value of σ_1^2 in H_1, the power of the scan statistics, based on a training sample approach and conditioning on the sufficient statistic approach, can be evaluated via the following algorithm:

1. Run N simulations of M independent observations under the specified alternative hypothesis. For the training sample method, simulate n_0 observations for the training sample in each run.

2. For each run of the simulation, compute the observed values of the scan statistics $S^*_{m,M}$ or $S^{**}_{m,M}$ given in equation (3.3) or (3.6), respectively.

3. Reject H_0 if the observed value of the scan statistic based on Monte Carlo

simulation under H_0 exceeds the critical value corresponding to the specified significance level α. The power equals the proportion of rejections out of the N simulation runs under H_1.

The power of the parametric bootstrap test can be evaluated as follows:

1. Run N simulations of M independent observations under the specified alternative hypothesis.

2. For the ith, $i = 1, \ldots, N$, simulated data sequence, compute the bootstrap P-value, denote by $\hat{p}_i$, by the bootstrap testing procedure given in (3.8).

3. Estimate the power by the proportion of these bootstrap P-values that are less than α :

$$\hat{\beta}_B = \frac{\#\{\hat{p}_i < \alpha; i = 1, \ldots, N\}}{N}.$$

Furthermore, a nice algorithm to produce a conservative confidence interval for the bootstrap test power, [1] given the length and coverage probability of the interval can be found in [12], which is already implemented in the R-package *simctest*, available on CRAN. One can refer to their paper for details if interested in getting a confidence intervals of the power.

In Section 3.5, Tables 3.1 and 3.2, for selected values of the parameters, we evaluate the power for the three fixed window scan statistics. From the numerical

[1] in fact, this algorithm works for almost any Monte Carlo type hypothesis test method

results, it is evident that the scan statistic based on conditioning on the sufficient statistic for σ_0^2 outperforms the other two scan statistics.

3.3 A Conditional Multiple Window Scan Statistic

A shortcoming for a fixed window scan statistic arises from the fact that in practice one usually does not know the true window size m, where a change in variance has occurred. Using a fixed window scan statistic, with an incorrect size for the moving window, will result in loss of power. One approach to address this problem, is to employ a multiple window scan statistic ([6,17,40,41]). Based on numerical results in Section 3.5, Tables 3.1 and 3.2, for fixed window scan statistics discussed in Section 3.2, we have concluded that the scan statistic conditional on the sufficient statistic for σ_0^2 is superior to the other two fixed window scan statistics. Hence, only the multiple window scan statistic via the conditioning on the sufficient statistic for σ_0^2 will be discussed below.

For $n \geq 2$, let $2 \leq m_1 < m_2 < \ldots < m_n \leq M/4$ be the n sliding windows chosen by the experimenter. For the transformed data sequence $\{X_1^{**}, ..., X_M^{**}\}$, defined in (3.5), the corresponding fixed window scan statistics, $S_{m_1,M}^{**}, \ldots, S_{m_n,M}^{**}$, are given in equation (3.6). For $1 \leq j \leq n$, let s_j,be the observed value of $S_{m_j,M}^{**}$ and $p_j = P(S_{m_j,M}^{**} > s_j | R^2 = r^2, H_0)$ its associated p-value, respectively. For testing H_0 vs. H_1, we propose to employ the following minimum P-value statistic,

denoted by P_{min}, defined as:

$$P_{min} = \min\{p_j; 1 \leq j \leq n\}. \tag{3.9}$$

This P_{min} is referred to as a *conditional multiple window scan statistic*. Note that, in the context of multiple testing, P_{min} can be viewed as a bootstrap test statistic ([8], Sec. 4.4.3). For the problem at hand, the null hypothesis is rejected if the observed value of P_{min} falls below a specified critical value. Since the exact distribution of the P_{min} statistic is unknown, for a given significant level α, the critical value p_α :

$$P_{H_0}(P_{min} < p_\alpha) = \alpha, \tag{3.10}$$

has to be computed by a Monte Carlo simulation. In Section 3.5, Tables 3.3 - 3.6, for selected values of the parameters, we evaluate the power for the P_{min} statistic.

While employing P_{min} to test H_0 vs. H_1, one can obtain an estimate of the window size where a change in variance has occurred, $\hat{m}$, from the window size corresponding to the observed value of P_{min}. Moreover, one can estimate the starting location of the window with the change of variance, $\hat{a}$, via the location which maximizes the moving sum squares with the fixed window size $\hat{m}$. We discuss our findings briefly in Section 3.5.

3.4 A Conditional Variable Window Scan Statistics

An alternative test statistic for the testing problem, outlined in Section 3.2, can be derived via the generalized likelihood ratio method, following the approach in

[32] and [29]. In our case, we derive a conditional generalized likelihood ratio test (GLRT), that is based on conditioning on the total sum of squares of the whole data sequence, $\sum_{k=1}^{M} X_k^2 = R^2$, and the sum of squares of the partial data, $\{X_a, \ldots, X_{a+m-1}\}$ corresponding to a specified alternative, $\sum_{k=a}^{a+m-1} X_k^2 = r^2$, where $3 \leq m \leq M/4$. Therefore, under H_0, $(X_1, X_2, \ldots, X_M)$, conditional on R, has a joint the uniform distribution on the $(M-1)$ sphere with radius R. Moreover, under H_1, conditional on R and r, $(X_1, \ldots, X_{a-1}, X_{a+m}, \ldots, X_M)$ jointly follow a uniform distribution on the $(M-m-1)$ sphere with radius $\sqrt{R^2-r^2}$ and are independent of $(X_a, \ldots, X_{a+m-1})$, where the latter jointly follow the uniform distribution on $(m-1)$ sphere with radius r. Hence, for the problem at hand, the conditional GLRT is given by:

$$
\begin{aligned}
\Lambda =& \frac{sup_{\Theta_1} f(x_1, \ldots, x_M \mid R, r)}{sup_{\Theta_0} f(x_1, \ldots, x_M \mid R, r)} \\
=& \frac{sup_{\Theta_1} \left\{ \frac{1}{SS_{m-1}(r)} \times \frac{1}{SS_{M-m-1}(\sqrt{R^2-r^2})} \right\}}{sup_{\Theta_0} \left\{ \frac{1}{SS_{M-1}(R)} \right\}} \\
=& \frac{sup_{\Theta_1} \left\{ \frac{1}{m\pi^{m/2} r^{m-1}/\Gamma(m/2+1)} \times \frac{1}{(M-m)\pi^{(M-m)/2}(R^2-r^2)^{(M-m-1)/2}/\Gamma(M/2-m/2+1)} \right\}}{sup_{\Theta_0} \left\{ \frac{1}{M\pi^{M/2} R^{M-1}/\Gamma(M/2+1)} \right\}},
\end{aligned}
\tag{3.11}
$$

where Θ_0 and Θ_1 represent the parameter spaces under H_0 and H_1, respectively; $f(x_1, \ldots, x_M \mid R, r)$ is the joint density of $X_1, \ldots, x_M$ conditional on R and r under respective hypotheses; the function $SS_N(K) = N\pi^{N/2} K^{N-1}/\Gamma(N/2+1)$

gives the surface area of the $(N-1)$ sphere with radius K; and for $\alpha > 0$, $\Gamma(\alpha) = \int_0^\infty \exp(-x)x^{\alpha-1}dx$ is the gamma function.

After routine derivations, it follows from equation (3.11), that

$$\begin{aligned}\Lambda =& \Lambda(m, a \mid R, r) \\ \propto& sup_{\Theta_1} \frac{B(m/2,(M-m)/2)}{(r^2/R^2)^{(m-1)/2}(1-r^2/R^2)^{(M-m-1)/2}} \\ \propto& sup_{\{m,a\}} \frac{B(m/2,(M-m)/2)}{(Y^{**}_{a,m})^{(m-1)/2}(1-Y^{**}_{a,m})^{(M-m-1)/2}},\end{aligned} \tag{3.12}$$

where $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ is the beta function and $Y^{**}_{a,m}$, as defined in (3.7), is the moving sum squares for the transformed data $\{X_i^{**} = X_i/R; 1 \le i \le M\}$, defined in (3.5).

Note that, the final representation of the conditional GLRT statistic depends only on the joint distribution of $\{X_i^{**} = X_i/R; 1 \le i \le M\}$, which under H_0, does not depend on the unknown value of σ_0, and under H_1, depends only on σ_1/σ_0. Moreover, for a fixed window size m, the function $g(Y^{**}) = (Y^{**})^{(m-1)/2}(1 - Y^{**})^{(M-m-1)/2}$ is decreasing in Y^{**} under H_1. Therefore, the conditional GLRT, Λ is increasing in Y^{**} Therefore, for a known fixed value of m, the conditional GLRT will be the same as our fixed window scan statistic, conditional on the sufficient statistic for σ_0^2, discussed in Section 3.2. When the window size m is unknown, equation (3.12) leads us to the following algorithm for implementing the conditional GLRT:

- Transform the data sequence $X_1, \ldots, X_M$ into $X_1^{**}, \ldots, X_M^{**}$ as defined in

(3.5)

- For each $3 \leq m \leq M/4$, execute the following steps.
  - Compute $Y^{**}_{a,m} = \sum_{i=a}^{a+m-1}(X^{**}_i)^2$ for all a, where $1 \leq a \leq M - m + 1$.
  - Find $\max\{Y^{**}_{a,m}; 1 \leq a \leq M - m + 1\}$, record it as $Y^{***}(m)$ and the corresponding a record as $a^*(m)$.
  - Substitute $Y^{***}(m)$ in equation (3.12) and denote it by $LR(m)$.
- Find $\max_{3\leq m\leq m/4}\{LR(m)\}$, record it as LR^* and record the corresponding m as m^*. Then LR^* is the value of the conditional GLRT statistic; m^* is the most likely window size where a possible upward local change in variance has occurred and $a^*(m^*)$ is the most likely starting location for the local change.

We refer to the conditional GLRT statistic, Λ, as a *conditional variable window scan statistic*. The associated critical value or p-value for Λ can be obtained by performing N simulation runs, each having M i.i.d. $N(0,1)$ random variables, and then repeating the above steps for each of the simulated M-sequences. When M is large, to avoid possible numerical underflow problems, one needs to consider the $log(\Lambda)$ statistic.

In Section 3.5, Tables 3.3-3.6, for selected values of the parameters, we present a power comparisons between the conditional multiple window scan statistic P_{min} and the conditional variable variable window scan statistic Λ. In Tables

3.5-3.6, the power of P_{min} and Λ is compared to the power of the conditional fixed window scan statistic $S^{**}_{m,M}$.

3.5 Numerical Results

In this section, numerical results are presented to evaluate, for selected values of the parameters, the power of scan statistics discussed in Sections 3.2-3.4. The power of the scan statistics is evaluated based on a simulation with $10,000$ trials, employing algorithms that have been presented in Sections 3.2-3.4. For each scan statistic investigated in this article, we have simulated a sequence of independent normal observations with $\mu = 0$ and $\sigma_0 = 1$ and length $M = 100$ or $M = 1000$, respectively, with selected window sizes specified in the tables. For each simulated data to be scanned, without loss of generality, the window of length m where the local change of variance occurs starts at the 11th observation. The window size for the local change of variance is: $m = 10$ in Tables 3.1-3.4, and $m = 7$ in Tables 3.5-3.6. The change in the local variance in the specified window is an upward shift of σ , from $\sigma_0 = 1$ to σ_1, as listed in the tables.

In Tables 3.1-3.2, for selected values of the parameters, we evaluate the performance, via power comparison, of fixed window scan statistic introduced in Section 3.2. In Tables 3.1-3.2, the power is evaluated for fixed window scan statistics based on the training sample, bootstrap and conditioning on the sufficient statistic approaches. In Table 3.1, the power is also evaluated for the double

bootstrap method. For the training sample approach, $n_0 = 20$ has been used for the simulated training sample size. For the parametric bootstrap and double bootstrap testing methods, the bootstrap sample size $B = 1500$ has been used. Based on the numerical results in Tables 3.1-3.2, it is evident that the fixed window scan statistic based on the conditioning on the sufficient statistics approach performs best. From the numerical results in Table 3.1, one can see that the scan statistic based on double bootstrap testing method performs as well as the scan statistic based on the conditioning on the sufficient statistics approach. For $\alpha = .01$ or a small local shift in variance, the fixed window scan statistic based on the training sample approach and the bootstrap approach have a lower power. Hence for the problem at hand, the training sample approach or the single bootstrap approach are not recommended. The fixed window scan statistic based on the double bootstrap approach provides significant improvement of power over the single bootstrap approach. However, it is computationally too intensive for computing the power for $M = 1000$. Therefore, considering both the performance and computational efficiency, the use of the fixed window scan statistic based on the conditional approach is recommended. From the numerical results in Tables 3.1-3.2, one can also observe that, for each method, there is a loss in power when the fixed scanning window size is far away from the size of the window where a change in the variance has occurred (here the size of that window is $m = 10$).

In Tables 3.3-3.4, for selected values of the parameters, we evaluate the

power for the conditional multiple window scan statistic, P_{min}, and the variable window scan statistic based on the conditional $GLRT$, that have discussed in Section 3.3 and 3.4. In both tables, the size of the window where a change in the variance has occurred ($m = 10$) is one of the windows considered by the P_{min} statistic. From the numerical results in Tables 3.3-3.4, one observes that both scan statistics perform similarly. They both perform well when the change in the variance is moderate or large.

In Tables 3.5-3.6, for selected values of the parameters, we evaluate the power for five fixed window scan statistics, the conditional multiple window scan statistic, P_{min}, and the variable window scan statistic based on the conditional $GLRT$, when the size of the window where a change in the variance has occurred is $m = 7$ while the window sizes of the fixed window scan statics and the P_{min} statistics are either larger or smaller, as specified in these tables. From Tables 3.5-3.6, it is evident that the fixed window scan statistics perform poorly compared to the multiple and variable window scan statistics. The variable window scan statistic based on the conditional $GLRT$ has a slightly higher power than the P_{min} statistic. For a large sequence of observation, such as $M = 1000$ or larger, the computing time for P_{min} statistic is significantly faster and is more practical for use in practice. For shorter sequences the use of the variable window scan statistic is recommended.

Table 3.1: Power Comparisons for 1-dim Adjusted Fixed Window Scan Statistics, σ_0^2 Unknown, $M = 100$

		$\sigma_1/\sigma_0 = \sqrt{2}$			$\sigma_1/\sigma_0 = 2$			$\sigma_1/\sigma_0 = 3$		
m		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
5	T	0.0426	0.1298	0.2121	0.2838	0.4976	0.6039	0.7910	0.8958	0.9365
	B	0.0388	0.1192	0.1902	0.3621	0.5576	0.6519	0.8541	0.9391	0.9630
	D	0.0674	0.1644	0.2398	0.4526	0.6179	0.7010	0.8984	0.9549	0.9696
	C	0.0716	0.1693	0.2480	0.4612	0.6252	0.7009	0.9020	0.9527	0.9665
10	T	0.0413	0.1304	0.2125	0.2863	0.4932	0.6111	0.7955	0.9039	0.9407
	B	0.0380	0.1178	0.1944	0.4109	0.6069	0.6961	0.9093	0.9584	0.9739
	D	0.0817	0.1843	0.2723	0.5339	0.6874	0.7611	0.9436	0.9723	0.9821
	C	0.0869	0.1823	0.2701	0.5570	0.6844	0.7554	0.9448	0.9683	0.9786
15	T	0.0337	0.1177	0.1973	0.2165	0.4338	0.5516	0.7199	0.8674	0.9144
	B	0.0235	0.0908	0.1669	0.3107	0.5344	0.6448	0.8584	0.9407	0.9619
	D	0.0717	0.1760	0.2655	0.4824	0.6535	0.7301	0.9267	0.9630	0.9786
	C	0.0749	0.1662	0.2556	0.5011	0.6524	0.7287	0.9302	0.9603	0.9734
20	T	0.0307	0.1084	0.1832	0.1768	0.3864	0.5047	0.6618	0.8323	0.8890
	B	0.0113	0.0668	0.1369	0.2176	0.4512	0.5773	0.7900	0.9129	0.9465
	D	0.0598	0.1586	0.2417	0.4286	0.6076	0.6902	0.9035	0.9528	0.9710
	C	0.0586	0.1509	0.2317	0.4377	0.6098	0.6902	0.9065	0.9493	0.9652
25	T	0.0292	0.0990	0.1758	0.1493	0.3400	0.4646	0.6057	0.7919	0.8654
	B	0.0050	0.0409	0.0983	0.1291	0.3483	0.4884	0.6860	0.8665	0.9166
	D	0.0438	0.1358	0.2096	0.3556	0.5417	0.6344	0.8693	0.9334	0.9568
	C	0.0437	0.1316	0.2107	0.3646	0.5511	0.6411	0.8721	0.9353	0.9549

Table 3.2: Power Comparisons for 1-dim Adjusted Fixed Window Scan Statistics, σ_0^2 Unknown, $M = 1000$

		$\sigma_1/\sigma_0 = \sqrt{2}$			$\sigma_1/\sigma_0 = 2$			$\sigma_1/\sigma_0 = 3$		
m		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
5	T	0.0180	0.0721	0.1287	0.1465	0.2971	0.3979	0.6191	0.7836	0.8478
	B	0.0373	0.1062	0.1712	0.4056	0.5412	0.6122	0.8989	0.9385	0.9541
	C	0.0419	0.1094	0.1841	0.4250	0.5466	0.6230	0.9036	0.9402	0.9567
10	T	0.0202	0.0732	0.1302	0.1499	0.3090	0.4139	0.6470	0.8048	0.8632
	B	0.0441	0.1172	0.1803	0.4721	0.5920	0.6600	0.9309	0.9575	0.9678
	C	0.0487	0.1254	0.1917	0.4848	0.6016	0.6692	0.9343	0.9587	0.9692
25	T	0.0138	0.0620	0.1142	0.0603	0.1781	0.2699	0.3880	0.6178	0.7158
	B	0.0240	0.0772	0.1380	0.3301	0.4647	0.5405	0.8760	0.9213	0.9374
	C	0.0272	0.0961	0.1586	0.3415	0.4912	0.5649	0.8806	0.9271	0.9418
50	T	0.0113	0.0596	0.1052	0.0287	0.1102	0.1815	0.1965	0.4112	0.5293
	B	0.0103	0.0515	0.0964	0.1774	0.3018	0.3837	0.7649	0.8429	0.8730
	C	0.0169	0.0648	0.1176	0.2051	0.3325	0.4107	0.7895	0.8560	0.8846
100	T	0.0110	0.0541	0.1001	0.0174	0.0743	0.1323	0.0822	0.2340	0.3380
	B	0.0065	0.0306	0.0665	0.0663	0.1474	0.2194	0.5703	0.6884	0.7491
	C	0.0106	0.0499	0.1004	0.0882	0.1891	0.2681	0.6101	0.7268	0.7804

Table 3.3: Power Comparisons for 1-dim Conditional Multiple and Variable Window Scan Statistics, σ_0^2 Unknown, $M = 100$

	$\sigma_1/\sigma_0 = \sqrt{2}$			$\sigma_1/\sigma_0 = 2$			$\sigma_1/\sigma_0 = 3$		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
MW	0.0737	0.1795	0.2581	0.5130	0.6707	0.7368	0.9308	0.9645	0.9742
VW	0.0778	0.1784	0.2565	0.5216	0.6682	0.7360	0.9330	0.9655	0.9750

Note: the true cluster size in simulated data is $m = 10$, while the MW approach window sizes are {5, 10 ,15, 20, 25} and the range of window sizes considered for VW approach is 3 to $M/4$.

Table 3.4: Power Comparisons for 1-dim Conditional Multiple and Variable Window Scan Statistics, σ_0^2 Unknown, $M = 1000$

	$\sigma_1/\sigma_0 = \sqrt{2}$			$\sigma_1/\sigma_0 = 2$			$\sigma_1/\sigma_0 = 3$		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
MW	0.0393	0.1059	0.1737	0.4461	0.5596	0.6262	0.9226	0.9484	0.9607
VW	0.0395	0.1114	0.1821	0.4472	0.5708	0.6420	0.9219	0.9502	0.9635

Note: the true cluster size in simulated data is $m = 10$, while the MW approach window sizes are {5, 10, 25, 50, 100} and the range of window sizes considered for VW approach is 3 to $M/4$.

Table 3.5: Power Comparisons for 1-dim Conditional Scan Statistics When The True Cluster Size is Not Captured in Multiple Window Sizes, σ_0^2 Unknown, $M = 100$

	$\sigma_1/\sigma_0 = \sqrt{2}$			$\sigma_1/\sigma_0 = 2$			$\sigma_1/\sigma_0 = 3$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
5	0.0597	0.1439	0.2219	0.3740	0.5237	0.6047	0.8246	0.8894	0.9189
10	0.0548	0.1425	0.2142	0.3686	0.5238	0.6065	0.8229	0.8911	0.9176
15	0.0450	0.1234	0.1961	0.3123	0.4744	0.5627	0.7818	0.8652	0.8980
20	0.0366	0.1023	0.1699	0.2591	0.4084	0.5066	0.7310	0.8256	0.8698
25	0.0273	0.0897	0.1511	0.2072	0.3602	0.4525	0.6776	0.7903	0.8378
MW	0.0517	0.1337	0.2064	0.3540	0.5086	0.5941	0.8184	0.8857	0.9150
VW	0.0592	0.1422	0.2181	0.3801	0.5230	0.6124	0.8334	0.8930	0.9197

Note: the true cluster size in data is $m = 7$, which is not among the MW approach window sizes {5, 10, 15, 20, 25}. The range of window sizes considered for VW approach is 3 to $M/4$.

Table 3.6: Power Comparisons for 1-dim Conditional Scan Statistics When The True Cluster Size is Not Captured in Multiple Window Sizes, σ_0^2 Unknown, $M = 1000$

	$\sigma_1/\sigma_0 = \sqrt{2}$			$\sigma_1/\sigma_0 = 2$			$\sigma_1/\sigma_0 = 3$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
5	0.0336	0.0906	0.1559	0.3028	0.4230	0.4959	0.8136	0.8644	0.8905
10	0.0296	0.0916	0.1532	0.2856	0.4135	0.4918	0.8052	0.8651	0.8940
25	0.0211	0.0694	0.1233	0.1908	0.2924	0.3682	0.7095	0.7853	0.8275
50	0.0141	0.0566	0.1062	0.1013	0.1873	0.2609	0.5641	0.6619	0.7176
100	0.0111	0.0525	0.1047	0.0460	0.1157	0.1804	0.3942	0.5174	0.5835
MW	0.0292	0.0835	0.1431	0.2785	0.3919	0.4634	0.8021	0.8530	0.8782
VW	0.0326	0.0913	0.1488	0.2967	0.4149	0.4909	0.8113	0.8643	0.8895

Note: the true cluster size in data is $m = 7$, which is not among the MW approach window sizes {5, 10, 25, 50, 100}. The range of window sizes considered for VW approach is 3 to $M/4$.

3.6 Concluding Remarks

In this chapter we investigated the performance of fixed, multiple and variable window scan statistics in detecting a local change in variance for a sequence of i.i.d normal observations, when the population variance of the underlying normal distribution is unknown. When the size of the window where a local change in variance has occurred is known, the fixed window scan statistic based on the conditioning on the sufficient statistic approach and the parametric double bootstrap test approach performed well, when the shift in variance was moderate or large. For a large sequence of observations the implementation of the scan statistic based on the double bootstrap approach is computationally impractical. Both multiple and variable window scan statistics investigated in Section 3.3 and 3.4 performed well, when the shift in variance was moderate or large. The power of the variable window scan statistic was slightly higher, when the size of window in which the local change in variance has occurred was not included as one of the windows for the multiple window scan statistics. For a large sequence of observations, the implementation of the multiple window scan statistic is much faster than the variable window scan statistic. Moreover, both the multiple and variable window scan statistics can be used to estimate the window size and the approximate location where a local change in variance has occurred. For example, in Table 3.3 based on a simulation with $10,000$ trials, when the window size where a change has occurred is $m = 10$, starting at the 11th observation and $\sigma_1 = 3$, the average

estimated window size is 7.59 for the P_{min} statistic and 8.76 for the conditional $GLRT$, while the average estimated starting location of the window is 12.48 for the P_{min} statistic and 11.85 for the conditional $GLRT$. Additional research is needed to investigate effective methods to estimate the location and size of the window where a change has occurred.

Chapter 4

Scan Statistics for Two-dimensional Normal Data

4.1 Introduction

Scan statistics for detecting a local change in the population mean for integer valued observations in a two dimensional rectangular region were introduced in [5]. Since then scan statistics for two dimensional data have been of great interest in the scientific literature. Similar to one dimensional case, most of the research in the area of two dimensional scan statistics has been focused on detecting a local change in the mean of the observed data. For normally distributed two dimensional data, [39] investigated the performance of fixed and multiple window scan statistics for detecting a local change in the mean.

In this dissertation, Chapter 2 and 3 introduced scan statistics for detecting a local change in population variance for a sequence of normal observations. In this chapter we investigate the performance of scan statistics for detecting a local change in variance for normal observations in a two dimensional rectangular region.

The chapter is organized as follows. In Section 4.2, we introduce two fixed window scan statistics based on moving sums of squares of observations in rectangular windows. The first scan statistic is for the case when the population variance under the null hypotheses of no local change is known. For this scan statistic we present two approximations for its distribution, that facilitate its implementation for a specified significance level. For the case when the population variance under the null hypotheses of no local change is unknown, we condition on its sufficient statistic. We present a simulation algorithm to approximate the distribution of this conditional scan statistic. In Section 4.3, we investigate the performance of nonparametric bootstrap type multiple window scan statistics, for the case when the correct size of the rectangular window where a change in variance has occurred is unknown. Both known and unknown population variance cases are considered. We present simulation algorithms to implement these scan statistics. In Section 4.4, we derive variable window scan statistics via the generalized likelihood ratio tests approach. We present simulation algorithm to implement these variable window scan statistics. In Section 4.5, for selected values of the parameters, we present numerical results to evaluate the power of the scan statistics discussed in Sections 4.2 - 4.4. Concluding remarks are presented in Section 4.6.

4.2 Two Dimensional Fixed Window Scan Statistics

For $1 \leq i \leq M_1$ and $1 \leq j \leq M_2$, let $\{X_{ij}\}$ be independent and identically distributed (i.i.d) normal observations with mean μ and variance σ_0^2. We are interested in detecting an occurrence of a local change in variance, from σ_0^2 to σ_1^2, within a rectangular subregion of $m_1 \times m_2$ observations, in the observed data within the $M_1 \times M_2$ two dimensional rectangular region. In this article, we investigate the detection of a local upward shift in variance. One can modify the methods in this article to accommodate detection of a local downward shift or a two-sided shift. For $k = 1, 2$, let $2 \leq m_k \leq M_k/4$ be the pre-specified size of a two dimensional sliding window. A fixed window *scan statistic* for detecting a local change in variance, is defined by ([43])

$$S_{m_1,m_2}(M_1, M_2) = max\{Y_{i_1,i_2}(m_1, m_2); 1 \leq i_k \leq M_k - m_k + 1, k = 1, 2\}, \quad (4.1)$$

where for $1 \leq i_k \leq M_k - m_k + 1, k = 1, 2$,

$$Y_{i_1,i_2}(m_1, m_2) = \sum_{i=i_1}^{i_1+m_1-1} \sum_{i=i_2}^{i_2+m_2-1} X_{ij}^2 \quad (4.2)$$

are the moving sums of squares in the $m_1 \times m_2$ rectangular grid of the observed data with south west location (i_1, i_2). To simplify the presentation of the results in this article we will assume that $M_1 = M_2 = M$, $m_1 = m_2 = m$ and $M = Lm$, where $L \geq 3$ is an integer.

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

$$G_{m,t}(M) = P(S_{m,m}(M,M) \leq t)$$
$$= P(max\{Y_{i_1,i_2}(m_1,m_2); 1 \leq i_k \leq M_k - m_k + 1, k = 1,2\} \leq t), \quad (4.3)$$

be the cumulative distribution function of $S_{m,m}(M,M)$. Then,

$$P(S_{m,m}(M,M) > t) = 1 - G_{m,t}(M). \quad (4.4)$$

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

For detecting a potential occurrence of a local change in variance, we will be testing the null hypothesis: H_0: $X_{ij}, 1 \leq i,j \leq M$, are i.i.d. normal observations with mean μ and variance σ_0^2. The alternative hypothesis is: H_a: $X_{ij}, 1 \leq i,j \leq M$, are independent normal observations with mean μ, the $X'_{ij}s$ have variance $\sigma_1^2 > \sigma_0^2$, for $i,j \in R_{a_1,a_2}(m,m) = \{(i_1,i_2); a_k \leq i_1, i_2 \leq a_k + m + 1, k = 1,2\}$, where $1 \leq a_1, a_2 \leq M - m + 1$ are unknown coordinates of the southwest location of an $m \times m$ window, and variance σ_0^2 for $i,j \notin R_{a_1,a_2}(m,m)$. The restriction on the size of the rectangular window, $m \leq M/4$, is used to emphasize the interest in detecting a local change in variance. For our hypotheses testing problem, without loss of generality, one can always assume that $\mu = 0$. If $\mu \neq 0$, one can replace the $X'_{ij}s$ with the following recurrent residuals:

$$W_{ij} = \frac{[(i-1)M + j - 1]X_{ij} - \sum_{i_1=1}^{i-1}\sum_{i_2=1}^{j-1} X_{i_1 i_2}}{\sqrt{[(i-1)M + j][(i-1)M + j - 1]}}, (i-1)M + j \geq 2,$$

which are then iid normal random variables with mean 0 and variance σ_0^2 under H_0. ([39]). We first discuss the case when σ_0^2 is known and without loss of generality assume $\sigma_0^2 = 1$.

When the true window size m where a change in variance has occurred, is known, the generalized likelihood ratio test rejects our null hypothesis, in favor of the local change alternative hypothesis H_a, whenever $S_{m,m}$ exceeds a threshold value t, where t is determined by $P(S_{m,m} \geq t|H_0) = \alpha$, where α is the specified significance level. Hence, to implement our testing procedure we need to evaluate accurately $G(M)$, the joint distribution of the moving sum of squares.

Under H_0, the random variables $\{Y_{i_1,i_2}(m_1, m_2); 1 \leq i_k \leq M_k - m_k + 1, k = 1, 2\}$, are m^2 -dependent and have a joint multivariate chi-square distribution and marginal chi-square distributions with m^2 degrees of freedom with a certain joint covariance matrix. Since there are no algorithms for computing this multivariate chi-square distribution, a simulation has to be employed. When M is large, to expedite the computations, we present two approximations by [39] and [23].

The first approximation,

$$G(M) \approx \frac{P(S_{m,m}(m+1, M) \leq t)^{M-m}}{P(S_{m,m}(m, M) \leq t)^{M-m-1}}, \tag{4.5}$$

is based on a related approximation for detecting a local change in mean for normal data in ([39], Equation 9). This approximation for the distribution of the scan statistic for the $M \times M$ rectangular region is based on scanning only the first m and $m + 1$ rows of the region with a sliding window of size $m \times m$.

A second approximation for $G(M)$, is based on the method in [23] for detecting a local change in mean for i.i.d. integer valued observations in a two dimensional rectangular region. A nice feature of this approximation is that it provides a sharp error bound. This approximation is valid for our testing problem and is given by:

$$G(M) \approx (2Q_2 - Q_3)[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{-L+1}, \tag{4.6}$$

where Q_2 and Q_3 are approximated by:

$$Q_2 \approx (2Q_{22} - Q_{23})[1 + Q_{22} - Q_{23} + 2(Q_{22} - Q_{23})^2]^{-L+1}$$

and

$$Q_3 \approx (2Q_{32} - Q_{33})[1 + Q_{32} - Q_{33} + 2(Q_{32} - Q_{33})^2]^{-L+1},$$

where $Q_{ab} = P(S_{m,m}(am, bm) \leq t)$, for $a, b = 2, 3$. This approximation is valid when $1 - Q_{22} \leq 0.025$, $1 - Q_{32} \leq 0.025$ and $1 - Q_2 \leq 0.025$. The error bound is approximated by

$$E = E_{app} + E_{sim} \tag{4.7}$$

where E_{app} arises from the approximation process in (4.6), E_{sim} arises from the simulation process of Q_{22}, Q_{23}, Q_{32}, and Q_{33}, and can be evaluated by:

$$E_{app} \approx 3.3(L-1)^2[(1 - Q_{22})^2 + (1 - Q_{32})^2 + (L-1)(Q_{22} - Q_{23})^2]$$

and

$$E_{sim} \approx (L-1)^2 \times 1.96\sqrt{\frac{R_1 + R_2 - 2R_3 - (R_1 - R_2)^2}{N}},$$

where N is the number of simulation runs to evaluate Q_{ab} by generating replications for corresponding $S_{m,m}(am, bm)$ for $a, b = 2, 3$. To evaluate R_1, R_2 and R_3, denote by ϕ^i_{ab} the ith replication of $S_{m,m}(am, bm)$ for $a, b = 2, 3$ and $i = 1, 2, \ldots, N$. Let $I(.)$ denote the indicator function. Then

$$R_1 = \frac{1}{N}\sum_{i=1}^{N} I((\phi^i_{22} \leq t) \cap (\phi^i_{23} > t)),$$

$$R_2 = \frac{1}{N}\sum_{i=1}^{N} I((\phi^i_{32} \leq t) \cap (\phi^i_{33} > t))$$

and

$$R_3 = \frac{1}{N}\sum_{i=1}^{N} I((\phi^i_{22} \leq t) \cap (\phi^i_{23} > t)) \times I((\phi^i_{32} \leq t) \cap (\phi^i_{33} > t)).$$

Notice that, both approximations are also valid for i.i.d. observations from distributions other than normal. In Section 4.5, for selected values of the parameters, in Tables 4.1 and 4.2, we present numerical results for approximation (4.5), (4.6) and (4.7).

When the population variance σ_0^2 is unknown, one can employ the conditioning on the sufficient statistics for σ^2 approach to implement the fixed window scan statistic, as suggested in [42]. for the one dimensional case. For the problem at hand, we condition on the sufficient statistic $R^2 = \sum_{i=1}^{M}\sum_{j=1}^{M} X_{ij}^2$ for σ_0^2 to eliminate the unknown parameters of the null distribution of $S_{m,m}$. Under H_0, the distribution of the random vector $\{X_{ij}, 1 \leq i, j \leq M\}$, given $R^2 = R_0^2$, is uniform on a sphere of radius R_0. ([9], Chap. 12). Moreover, the random vector

$$\{X^*_{ij} = X_{ij}/R; 1 \leq i, j \leq M\}, \tag{4.8}$$

where $R = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{M} X_{ij}^2}$, has a joint uniform distribution on the (M^2-1) dimensional unit sphere. Consequently, we can define a scan statistic for the sequence of observations $\{X_{ij}^*; 1 \le i,j \le M\}$:

$$S_{m,m}^*(M,M) = max\{Y_{i_1,i_2}^*(m,m); 1 \le i_k \le M-m+1, k=1,2\}, \tag{4.9}$$

where

$$Y_{i_1,i_2}^*(m,m) = \sum_{i=i_1}^{i_1+m-1}\sum_{i=i_2}^{i_2+m-1} X_{ij}^{*2} = \frac{\sum_{i=i_1}^{i_1+m-1}\sum_{i=i_2}^{i_2+m-1} X_{ij}^2}{R^2}, \tag{4.10}$$

for $1 \le i_k \le M-m+1, k=1,2$. We abbreviate $S_{m,m}^*(M,M)$ to $S_{m,m}^*$.

We propose to employ this scan statistic conditional on $R^2 = R_0^2$ for testing H_0 when the variance σ_0^2 is unknown. Under H_0, the distribution of $S_{m,m}^*$ does not depend on any unknown parameters. Hence, for a given significance level α we can find the critical value t such that $P(S_{m,m}^* \ge t | R^2 = R_0^2, H_0) = \alpha$ and evaluate the conditional *P-value* of this scan statistic given by $P(S_{m,m}^*(M,M) \ge s | R^2 = R_0^2, H_0)$, where s is the observed value of $S_{m,m}^*$. These computations will be implemented via a Monte Carlo simulation that generates N sequences of data of $M \times M$ i.i.d. $N(0,1)$ observations, and then dividing each observation by R. The power of $S_{m,m}^*$ can be evaluated by the following algorithm:

1. Run N simulations of $M \times M$ independent observations under the specified alternative hypothesis.

2. For each run of the simulation, compute the observed values of the scan statistics $S^*_{m,m}$ given in equation (4.9).

3. Reject H_0 if the observed value of the scan statistic based on Monte Carlo simulation under H_0 exceeds the critical value corresponding to the specified significance level α or if the corresponding conditional *P-value* $< \alpha$. The power equals the proportion of rejections out of the N simulation runs under H_1.

In Section 4.5, for selected values of the parameters, in Tables 4.3 - 4.10, we evaluate the performance of fixed window scan statistics when σ_0^2 is assumed to be known and when σ_0^2 is unknown.

4.3 Two Dimensional Multiple Window Scan Statistics via the Minimum P-value Approach

The use of fixed window scan statistics is limited in practice as one usually does not know precise size of the $m \times m$ rectangular region where a change in the variance has occurred. Using a fixed window scan statistic, with an incorrect size for the moving window, will result in loss of power. One approach to address this problem, is to employ a *multiple window scan statistic* ([39,6,17,41,42]). When the size of the rectangular window, where a change in the variance has occurred is unknown, one can employ simultaneously a sequence of n fixed window scan statistics: $\{S_{m_1,m_1}, S_{m_2,m_2}, \ldots, S_{m_n,m_n}\}$, where the sizes of the rectangular

windows $2 \leq m_1 < m_2 < \ldots < m_n \leq M/4$ are chosen in advance by the experimenter.

We first discuss the case when σ_0^2 is known. For $1 \leq k \leq n$, let t_k be the observed value of S_{m_k,m_k} and $p_k = P(S_{m_k,m_k} > t_k|H_0)$ its associated *P-value*. To test H_0 vs. H_a, the minimum *P-value* statistic, P_{min}, is defined as follows:

$$P_{min} = \min\{p_k; 1 \leq k \leq n\}. \tag{4.11}$$

In the context of multiple testing, one can view the P_{min} statistic as nonparametric bootstrap test statistic ([8], Sec. 4.4.3). The null hypothesis is rejected if the observed value of P_{min} falls below a critical value corresponding to a specified significance level α. Since the exact distribution of the P_{min} statistic is unknown, for a given significant level α, the critical value p_α,

$$P_{H_0}(P_{min} < p_\alpha) = \alpha, \tag{4.12}$$

has to be evaluated by a Monte Carlo simulation. The following algorithm can be used to find the critical value p_α:

- Run N simulations with $M \times M$ i.i.d. $N(0,1)$ observations generated in each run of the simulation.

- For each run of the simulation, record the observed values of the fixed window scan statistics, $S_{m_1,m_1}, \ldots, S_{m_n,m_n}$, denoted by $t_1, \ldots, t_n$, respectively. For $1 \leq k \leq n$, compute the observed *P-value*, $p_k = P_{H_0}(S_{m_k,m_k} > t_k)$, from the simulations obtained in previous steps.

- For each run of the simulation, record the minimum value of $p_k, 1 \leq k \leq n$, as P_{min}.

- p_α will be the $(\alpha \times 100)$th percentile of the N P_{min}'s.

When the variance σ_0^2 is unknown, a *conditional* multiple window scan statistics P^*_{min} can be constructed similarly to P_{min}, based on the observed *P-values* of a sequence of n fixed window scan statistics: $\{S^*_{m_1,m_1}, S^*_{m_2,m_2}, \ldots, S^*_{m_n,m_n}\}$, where $S^*_{m_k,m_k}, 1 \leq k \leq n$, are defined in Section 4.2, Equation (4.9).

In Section 4.5, for selected values of the parameters, in Tables 4.3-4.10, we evaluate the power for the P_{min} and P^*_{min} statistics, for σ_0^2 known and unknown, respectively. Note that, while employing the P_{min} or P^*_{min} statistics, respectively, to test H_0 vs. H_1, one can obtain an estimate of the size of the rectangular window, denoted by $\hat{m}$, where a change in variance has occurred, corresponding to the observed value of P_{min} or P^*_{min}, respectively. Moreover, one can estimate the south-west location, (a, b), of the $\hat{m} \times \hat{m}$ rectangular region where the change in the variance has occurred. This estimate is based on south-west location that maximizes the moving sum squares within any fixed rectangular window of size $\hat{m} \times \hat{m}$. We discuss our findings briefly in Section 4.5, Table 4.11 and 4.12.

4.4 Two Dimensional Variable Window Scan Statistics via the Generalized Likelihood Ratio Method

When σ_0^2 is known, an alternative approach to the multiple window scan statistic P_{min}, is to derive a *variable window scan statistic* via the generalized likelihood ratio test (GLRT) principle, following the approach in [32] and [29]. When σ_0^2 is known, for our testing problem of detecting a local upward shift in variance, the generalized likelihood ratio test will reject H_0 in favor of H_1 for large values of

$$\Lambda = \frac{\sup_{\theta \epsilon \Theta_1} \prod_{i,j=1,\ldots,M} f_\theta(x_{ij})}{\sup_{\theta \epsilon \Theta_0} \prod_{i,j=1,\ldots,M} f_\theta(x_{ij})}, \tag{4.13}$$

where $f_\theta(x_{ij})$ is the probability density of the (i,j)th observation in the scanned area $\{X_{ij}\}$ and Θ_0 and Θ_1 are the parameter spaces for the null and alternative hypotheses, respectively. Suppose the true local change cluster starts at the (a,b)th observation with size $m \times m$, then this generalized likelihood ratio statistic can be expressed explicitly as:

$$\begin{aligned}
\Lambda &= \sup_{\Theta_1} \left(\frac{1}{\sigma_1}\right)^{m^2} exp\left(\frac{1}{2}\sum_{i=a}^{a+m-1}\sum_{j=b}^{b+m-1} X_{ij}^2 - \frac{1}{2\sigma_1^2}\sum_{i=a}^{a+m-1}\sum_{j=b}^{b+m-1} X_{ij}^2\right) \\
&= \sup_{\Theta_1} \left(\frac{1}{\sigma_1}\right)^{m^2} exp\left(\frac{1}{2}Y_{a,b}(m,m) - \frac{1}{2\sigma_1^2}Y_{a,b}(m,m)\right) \\
&= \sup_{a,b;m} \left(\frac{m^2}{Y_{a,b}(m,m)}\right)^{m^2/2} exp\left(\frac{1}{2}Y_{a,b}(m,m) - \frac{m^2}{2}\right),
\end{aligned} \tag{4.14}$$

where $Y_{a,b}(m,m) = \sum_{i=a}^{a+m-1}\sum_{j=b}^{b+m-1} X_{ij}^2$. The last step follows from the fact that for fixed and but arbitrary (a,b) and m, constrained by parameter space Θ_1, the

supremum is achieved at $\hat{\sigma}_1^2 = Y_{a,b}(m,m)/m^2 > 1$. Let

$$L_m(Y_{a,b}(m,m)) = \left(\frac{m^2}{Y_{a,b}(m,m)}\right)^{m^2/2} exp\left(\frac{1}{2}Y_{a,b}(m,m) - \frac{m^2}{2}\right). \qquad (4.15)$$

Regard $L_m(Y_{a,b}(m,m))$ as a function of $Y = Y_{a,b}(m,m)$, depending on (a,b) and m, For fixed but arbitrary m, this function is a convex function of Y and it is increasing in Y on Θ_1. Therefore, for fixed m, the supremum in (4.14) is achieved at the maximum value of Y. One can obtain a unique value of (a,b) that maximizes Y. It follows that, for a given scanning area of observations, one can get the location and size of the window that maximizes $L_m(Y_{a,b}(m,m))$. This maximum value of $L_m(Y_{a,b}(m,m))$ is the value of our variable window scan statistic based on the generalized likelihood ratio principle. For a given area of observations $\{X_{11},\ldots,X_{MM}\}$ in our testing problem, the algorithm presented below implements the search for the location and size of the window that maximizes $L_m(Y_{a,b}(m,m))$.

- For $3 \le m \le M/4$, execute the following steps.
  - Compute $Y_{a,b}(m,m) = \sum_{i=a}^{a+m-1}\sum_{j=b}^{b+m-1} X_{ij}^2$ for all (a,b), where $1 \le a,b \le M-m+1$.
  - Find $\max\{Y_{a,b}(m,m); 1 \le a,b \le M-m+1\}$ and record as $Y^*(m)$ and the corresponding (a,b) record as $(a^*(m), b^*(m))$.
  - Compute $L^*(m) = L_m(Y^*(m))$ by equation (4.15).

- Find $\max_{3\leq m\leq M/4}\{L^*(m)\}$, record it as Λ^* and record the corresponding m as m^*. Then Λ^* is the value of the variable window scan statistic; $m^*\times m^*$ is the most likely window size where a possible upward local change in variance has occurred and $(a^*(m^*), b^*(m^*))$ is the most likely starting position for the local change.

The *P-value* corresponding to the observed value of Λ^* can be found by a simulation algorithm as follows:

- Perform N simulations, each generating $M\times M$ i.i.d. $N(0,1)$ random variables.

- For each simulation, compute the observed value of Λ in the same way as Λ^* has been evaluated by the algorithm presented above.

- The *P-value* is equal to the proportion of the observed values of Λ's, based on the N simulations, that exceeds Λ^* for the data set

In Section 4.5, Tables 4.3 - 4.6 for selected values of the parameters, we compare the performance of the variable window scan statistic based on the generalized likelihood ratio principle with the fixed window scan statistics and the multiple window scan statistics via P_{min} proposed in Section 4.2 and 4.3.

When the population variance σ_0^2 is unknown, we derive a *conditional* variable window scan statistic by conditioning on both the total sum of squares of the

whole scanning area, $\sum_{i=1}^{M}\sum_{j=1}^{M} X_{ij}^2 = R^2$, and the sum of squares of the partial data, $\sum_{i=a}^{a+m-1}\sum_{j=b}^{b+m-1} X_{ij}^2 = r^2$, where $3 \leq m \leq M/4$, corresponding to a specified alternative. Therefore, under H_0, $(X_{11}, X_{12}, \ldots, X_{MM})$ conditional on R, has a joint uniform distribution on the (M^2-1) sphere with radius R. Moreover, under H_1, conditional on R and r, the observations in the rectangular window where a local change in variance has occurred, $\{X_{ij}; a \leq i \leq a+m-1, b \leq j \leq b+m-1\}$ jointly follow a uniform distribution on the m^2-1 sphere with radius r, and are independent of the rest of the observations $\{X_{ij}\}$, where the latter jointly follow a uniform distribution on the (M^2-m^2-1) sphere with radius $\sqrt{R^2-r^2}$. Hence, for the problem at hand, the *conditional* GLRT is given by:

$$
\begin{aligned}
\Lambda_c =& \frac{sup_{\Theta_1} f(x_{11}, \ldots, x_{MM} \mid R, r)}{sup_{\Theta_0} f(x_{11}, \ldots, x_{MM} \mid R, r)} \\
=& \frac{sup_{\Theta_1}\left\{\frac{1}{SS_{m^2-1}(r)} \times \frac{1}{SS_{M^2-m^2-1}(\sqrt{R^2-r^2})}\right\}}{sup_{\Theta_0}\left\{\frac{1}{SS_{M^2-1}(R)}\right\}} \\
=& \frac{sup_{\Theta_1}\left\{\frac{1}{m^2\pi^{m^2/2}r^{m^2-1}/\Gamma(m^2/2+1)} \times \frac{1}{(M^2-m^2)\pi^{(M^2-m^2)/2}(R^2-r^2)^{(M^2-m^2-1)/2}/\Gamma(M^2/2-m^2/2+1)}\right\}}{sup_{\Theta_0}\left\{\frac{1}{M^2\pi^{M^2/2}R^{M^2-1}/\Gamma(M^2/2+1)}\right\}},
\end{aligned}
\tag{4.16}
$$

where Θ_0 and Θ_1 represent the parameter spaces under H_0 and H_1, respectively; $f(x_{11}, \ldots, x_{MM} \mid R, r)$ is the joint density of $X_{11}, \ldots, X_{MM}$ conditional on R and r under respective hypotheses; the function $SS_N(K) = N\pi^{N/2}K^{N-1}/\Gamma(N/2+1)$ gives the surface area of the $(N-1)$ sphere with radius K; and for $\alpha > 0$,

$\Gamma(\alpha) = \int_0^\infty \exp(-x)x^{\alpha-1}dx$ is the gamma function.

After routine derivations, it follows from equation (4.16), that

$$\begin{aligned}\Lambda_c =& \Lambda(m, a, b \mid R, r) \\ \propto& sup_{\Theta_1} \frac{B(m^2/2, (M^2-m^2)/2)}{(r^2/R^2)^{(m^2-1)/2}(1-r^2/R^2)^{(M^2-m^2-1)/2}} \\ \propto& sup_{\{m,a,b\}} \frac{B(m^2/2, (M^2-m^2)/2)}{(Y^*_{a,b}(m,m))^{(m^2-1)/2}(1-Y^*_{a,b}(m,m))^{(M^2-m^2-1)/2}},\end{aligned} \tag{4.17}$$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ is the beta function and $Y^*_{a,b}(m,m)$, as defined in (4.10), is the moving sum squares for the transformed data $\{X^*_{ij} = X_{ij}/R; 1 \leq i, j \leq M\}$, defined in (4.8).

Note that, the final representation of the conditional GLRT statistic depends only on the joint distribution of $\{X^*_{ij} = X_i/R; 1 \leq i, j \leq M\}$, which under H_0, does not depend on the unknown value of σ_0, and under H_1, depends only on σ_1/σ_0. Moreover, for a fixed rectangular window of size $m \times m$, the function $g(Y^*) = (Y^*)^{(m^2-1)/2}(1-Y^*)^{(M^2-m^2-1)/2}$ is decreasing in Y^* under H_1. Therefore, the conditional GLRT, Λ_c is increasing in Y^*. Therefore, for a known fixed value of m, the conditional GLRT will be coincide with our fixed window scan statistic, conditional on the sufficient statistic for σ_0^2, discussed in Section 4.2. When the size of the rectangular window, m, where the change in the variance has occurred is unknown, equation (4.17) leads us to the following algorithm for implementing the conditional GLRT:

- Transform the data sequence $X_{11}, \ldots, X_{MM}$ into $X^*_{11}, \ldots, X^*_{MM}$ as defined

in (4.8)

- For each $3 \leq m \leq M/4$, execute the following steps.
  - Compute $Y^*_{a,b}(m,m) = \sum_{i=a}^{a+m-1} \sum_{j=b}^{b+m-1} (X^*_{ij})^2$ for all a, b where $1 \leq a, b \leq M - m + 1$.
  - Find $\max\{Y^*_{a,b}(m,m); 1 \leq a, b \leq M - m + 1\}$, record it as $Y^{**}(m)$ and the corresponding (a, b) record as $(a^*(m), b^*(m))$.
  - Substitute $Y^{**}(m)$ in equation (4.17) and denote it by $LR(m)$.
- Find $\max_{3 \leq m \leq m/4}\{LR(m)\}$, record it as LR^* and record the corresponding m as m^*. Then LR^* is the value of the conditional GLRT statistic; m^* is the most likely window size where a possible upward local change in variance has occurred and $(a^*(m^*), b^*(m^*))$ is the most likely starting location for the local change.

We refer to this conditional GLRT statistic, Λ_c, as a *conditional* variable window scan statistic. The associated *P-value* for Λ_c can be obtained by performing N simulation runs, each having $M \times M$ iid $N(0,1)$ random variables, and then repeating the above steps for each of the simulated $M \times M$-sequences. When M is large, to avoid possible numerical underflow problems, one needs to consider the $log(\Lambda_c)$ statistic.

In Section 4.5, Tables 4.7 - 4.10 for selected values of the parameters, we compare the performance of the conditional variable window scan statistic Λ_c with

the conditional fixed window scan statistics and the conditional multiple window scan statistics P^*_{min} proposed in Section 4.2 and 4.3, respectively.

4.5 Numerical Results

In Tables 4.1 - 4.2, for selected values of the parameters, we present numerical results of a simulation study to evaluate the accuracy of two approximations for $P(S_{m,m}(M,M) > t)$, given in Section 4.2, Equation (4.5) and (4.6), marked as Approximation 1 and 2, respectively. We also present in these tables the error bound for Approximation 2, given in Equation (4.7). We have used $N = 100,000$ trials of simulated i.i.d. $N(0,1)$ observations for all simulations to evaluate the two approximations and the error bound. From the numerical results presented in Tables 4.1 - 4.2, it is evident that the two approximations performed well. Based on this simulation study, Approximation 1 performed slightly better for large M.

In Tables 4.3 - 4.6, for selected values of the parameters and three local shifts in variance, represented by σ_1/σ_0 and assuming σ_0 is known, we present numerical results for the power of fixed window, multiple window (MW) and variable window (VW) scan statistics, based on a simulation with $N = 10,000$ trials of simulated i.i.d. $N(0,1)$ observations. In Tables 4.3 - 4.6, the local change of variance occurred in an $m_0 \times m_0$ sub-region with a south-west location being $(11,11)$, within the rectangular $M \times M$ region. For each simulated data, we employ five fixed window scan statistics with an $m \times m$ scanning window, specified in each

table. The multiple window scan statistic is based on all five sizes of the scanning windows. The variable window scan statistic is based on all rectangular scanning windows of size $3 \leq m \leq 30$. Note that in Tables 4.5 and 4.6, the actual window size where a local change has occurred is not among the selected window sizes for the fixed or multiple window scan statistics. In Tables 4.7 - 4.10, with the same simulation parameters, when σ_0^2 is unknown, we evaluate the power of the conditional fixed, multiple and variable window scan statistics. In Tables 4.9 - 4.10, the actual window size for a local change in variance is not among the selected rectangular window sizes for the conditional fixed or multiple window scan statistics.

From the numerical results presented in Tables 4.3 and 4.4, it is evident that the power is maximized by using a fixed window scan statistic with the correct window size where a change in variance has occurred. When the correct size of the rectangular window where a change in the variance has occurred is unknown, using a fixed window scan statistic with an incorrect window size most often will result in sizable loss of power. The power can be significantly enhanced by employing a multiple or a variable window scan statistic. Based on our power simulation study, when the correct window size is among the sizes selected by the multiple window scan statistic, the multiple and variable window scan statistics perform equally well. However, when the correct window size, where a change in variance has occurred, is not included in the multiple window scan statistic, the

variable window scan statistic outperforms the multiple window scan statistic, as it is shown by the numerical results in Tables 4.5 and 4.6. Similar conclusions can be drawn from the numerical results presented in Tables 4.7 - 4.10 for the conditional fixed, multiple and variable window scan statistics, for the case when σ_0^2 is unknown.

In Tables 4.11 - 4.12 for selected values of the parameters, we present simulation results for the estimated size of the rectangular $\widehat{m} \times \widehat{m}$ subregion and its estimated south-west location $(\widehat{a}, \widehat{b})$ within the $M \times \ M$ rectangular region, via multiple and variable window scan statistics, for the cases when σ_0^2 is known and unknown, respectively. In both tables the numerical results are based on $N = 10,000$ simulation trials, for each specified value of M and method, with a shift ratio of $\sigma_1/\sigma_0 = 1.75$ for the local change of variance. The numerical results suggest that the variable and conditional variable window scan statistics estimated quite accurately the unknown size of the rectangular window and its location, where a change in variance has occurred. It performed better than the multiple and conditional multiple window scan statistics.

Table 4.1: Approximations of $P(S_{m,m}(M, M) > t)$ for 2-dim Fixed Window Scan Statistics, $M = 100$, $m = 5$

t	65	66	67	68	69	70	71
Direct Simulation	0.0928	0.0692	0.0499	0.0358	0.0253	0.0173	0.0123
Approximation 1	0.0989	0.0732	0.0525	0.0387	0.0275	0.0153	0.0134
Approximation 2	0.0935	0.0820	0.0668	0.0565	0.0400	0.0332	0.0202
Error Bound	0.0404	0.0358	0.0320	0.0294	0.0246	0.0224	0.0174

Table 4.2: Approximations of $P(S_{m,m}(M, M) > t)$ for 2-dim Fixed Window Scan Statistics, $M = 250$, $m = 10$

t	175	176	177	178	179	180	181
Direct Simulation	0.0993	0.0805	0.0650	0.0517	0.0420	0.0335	0.0269
Approximation 1	0.1075	0.0944	0.0767	0.0632	0.0495	0.0333	0.0286
Approximation 2	0.0857	0.0691	0.0527	0.0360	0.0300	0.0144	0.0080
Error Bound	0.0475	0.0414	0.0352	0.0285	0.0257	0.0164	0.0115

Table 4.3: Power Comparison for the 2-dim Scan Statistics, σ_0^2 Known, $M = 100$, $m_0 = 10$

	$\sigma_1/\sigma_0 = 1.25$			$\sigma_1/\sigma_0 = 1.50$			$\sigma_1/\sigma_0 = 1.75$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.0971	0.2275	0.3138	0.7439	0.8684	0.9123	0.9913	0.9975	0.9987
10	0.2493	0.4077	0.4979	0.9601	0.9810	0.9883	0.9999	1.0000	1.0000
15	0.1614	0.3162	0.4221	0.8863	0.9458	0.9656	0.9983	0.9996	0.9997
20	0.1108	0.2431	0.3444	0.7781	0.8808	0.9196	0.9925	0.9973	0.9985
25	0.0634	0.1591	0.2561	0.5671	0.7339	0.8075	0.9624	0.9847	0.9913
MW	0.1981	0.3537	0.4503	0.9353	0.9720	0.9814	0.9998	1.0000	1.0000
VW	0.2049	0.3531	0.4451	0.9372	0.9710	0.9805	0.9998	0.9999	0.9999

Table 4.4: Power Comparison for the 2-dim Scan Statistics, σ_0^2 Known, $M = 250$, $m_0 = 10$

	$\sigma_1/\sigma_0 = 1.25$			$\sigma_1/\sigma_0 = 1.50$			$\sigma_1/\sigma_0 = 1.75$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.0506	0.1306	0.2046	0.6027	0.7466	0.8091	0.9745	0.9901	0.9942
10	0.1409	0.2654	0.3470	0.9148	0.9547	0.9659	0.9994	0.9998	0.9998
15	0.0748	0.1759	0.2608	0.7825	0.8716	0.9073	0.9955	0.9984	0.9992
20	0.0480	0.1215	0.1976	0.6132	0.7498	0.8087	0.9800	0.9924	0.9953
25	0.0263	0.0864	0.1487	0.3755	0.5271	0.6149	0.9030	0.9494	0.9669
MW	0.1046	0.2094	0.2947	0.8781	0.9300	0.9498	0.9993	0.9996	0.9997
VW	0.1167	0.2211	0.3044	0.8838	0.9315	0.9488	0.9992	0.9995	0.9996

Table 4.5: Power Comparison for the 2-dim Scan Statistics, σ_0^2 Known, $M = 100$, $m_0 = 7$

	$\sigma_1/\sigma_0 = 1.25$			$\sigma_1/\sigma_0 = 1.50$			$\sigma_1/\sigma_0 = 1.75$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.0396	0.1154	0.1851	0.4388	0.6000	0.6674	0.8792	0.9346	0.9546
10	0.0395	0.1261	0.2009	0.4462	0.6068	0.6863	0.8893	0.9403	0.9583
15	0.0260	0.0987	0.1762	0.2663	0.4435	0.5490	0.7451	0.8537	0.8963
20	0.0197	0.0824	0.1443	0.1482	0.2925	0.3915	0.5476	0.6964	0.7736
25	0.0167	0.0682	0.1327	0.0798	0.1915	0.2910	0.3482	0.5189	0.6234
MW	0.0399	0.1156	0.1951	0.4376	0.5904	0.6746	0.8851	0.9365	0.9584
VW	0.0469	0.1313	0.2052	0.5152	0.6537	0.7229	0.9241	0.9601	0.9710

Table 4.6: Power Comparison for the 2-dim Scan Statistics, σ_0^2 Known, $M = 250$, $m_0 = 7$

	$\sigma_1/\sigma_0 = 1.25$			$\sigma_1/\sigma_0 = 1.50$			$\sigma_1/\sigma_0 = 1.75$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.0223	0.0786	0.1374	0.3221	0.4484	0.5304	0.7956	0.8722	0.9021
10	0.0238	0.0825	0.1416	0.3111	0.4520	0.5309	0.8085	0.8833	0.9088
15	0.0143	0.0655	0.1264	0.1603	0.2811	0.3690	0.5958	0.7221	0.7849
20	0.0127	0.0587	0.1109	0.0741	0.1631	0.2422	0.3530	0.5025	0.5869
25	0.0141	0.0540	0.1070	0.0358	0.1000	0.1695	0.1927	0.3198	0.4070
MW	0.0213	0.0740	0.1360	0.3043	0.4421	0.5214	0.7999	0.8769	0.9061
VW	0.0282	0.0880	0.1491	0.3903	0.5152	0.5870	0.8696	0.9209	0.9373

Table 4.7: Power Comparison for the 2-dim Scan Statistics, σ_0^2 Unknown, $M = 100$, $m_0 = 10$

	$\sigma_1/\sigma_0 = 1.25$			$\sigma_1/\sigma_0 = 1.50$			$\sigma_1/\sigma_0 = 1.75$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.0915	0.2179	0.3038	0.7251	0.8588	0.9045	0.9892	0.9971	0.9988
10	0.2366	0.3971	0.4807	0.9550	0.9801	0.9862	0.9999	0.9999	1.0000
15	0.1617	0.2996	0.3966	0.8810	0.9394	0.9590	0.9985	0.9995	0.9998
20	0.1032	0.2262	0.3162	0.7608	0.8646	0.9049	0.9906	0.9966	0.9980
25	0.0568	0.1429	0.2180	0.5208	0.6882	0.7686	0.9516	0.9815	0.9876
MW	0.1908	0.3398	0.4313	0.9306	0.9681	0.9779	0.9999	0.9999	1.0000
VW	0.1918	0.3328	0.4220	0.9309	0.9664	0.9762	0.9998	0.9999	0.9999

Table 4.8: Power Comparison for the 2-dim Scan Statistics, σ_0^2 Unknown, $M = 250$, $m_0 = 10$

	$\sigma_1/\sigma_0 = 1.25$			$\sigma_1/\sigma_0 = 1.50$			$\sigma_1/\sigma_0 = 1.75$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.0506	0.1294	0.2029	0.6008	0.7446	0.8076	0.9741	0.9899	0.9939
10	0.1427	0.2639	0.3402	0.9151	0.9530	0.9654	0.9995	0.9997	0.9998
15	0.0737	0.1737	0.2553	0.7770	0.8696	0.9042	0.9949	0.9982	0.9991
20	0.0447	0.1174	0.1895	0.6005	0.7444	0.8051	0.9790	0.9916	0.9947
25	0.0260	0.0809	0.1419	0.3707	0.5181	0.6061	0.9002	0.9468	0.9652
MW	0.1020	0.2099	0.2931	0.8750	0.9292	0.9494	0.9992	0.9996	0.9997
VW	0.1121	0.2141	0.2910	0.8780	0.9290	0.9454	0.9990	0.9995	0.9996

Table 4.9: Power Comparison for the 2-dim Scan Statistics, σ_0^2 Unknown, $M = 100$, $m_0 = 7$

	$\sigma_1/\sigma_0 = 1.25$			$\sigma_1/\sigma_0 = 1.50$			$\sigma_1/\sigma_0 = 1.75$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.0390	0.1138	0.1827	0.4296	0.5910	0.6660	0.8733	0.9333	0.9512
10	0.0381	0.1207	0.1914	0.4301	0.6002	0.6720	0.8795	0.9373	0.9556
15	0.0276	0.0948	0.1652	0.2650	0.4277	0.5299	0.7406	0.8447	0.8900
20	0.0184	0.0769	0.1320	0.1384	0.2727	0.3627	0.5226	0.6818	0.7525
25	0.0144	0.0625	0.1174	0.0675	0.1686	0.2526	0.3158	0.4815	0.5757
MW	0.0364	0.1116	0.1799	0.4249	0.5779	0.6561	0.8766	0.9346	0.9544
VW	0.0458	0.1262	0.1975	0.5094	0.6480	0.7137	0.9216	0.9600	0.9694

Table 4.10: Power Comparison for the 2-dim Scan Statistics, σ_0^2 Unknown, $M = 250$, $m_0 = 7$

	$\sigma_1/\sigma_0 = 1.25$			$\sigma_1/\sigma_0 = 1.50$			$\sigma_1/\sigma_0 = 1.75$		
m	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
5	0.0223	0.0785	0.1378	0.3219	0.4457	0.5284	0.7955	0.8718	0.9016
10	0.0237	0.0808	0.1356	0.3120	0.4484	0.5251	0.8083	0.8814	0.9082
15	0.0138	0.0635	0.1251	0.1559	0.2771	0.3630	0.5885	0.7203	0.7805
20	0.0118	0.0574	0.1074	0.0705	0.1587	0.2361	0.3431	0.4973	0.5802
25	0.0125	0.0532	0.1047	0.0365	0.0980	0.1632	0.1902	0.3094	0.3981
MW	0.0201	0.0739	0.1376	0.3004	0.4449	0.5212	0.7976	0.8772	0.9067
VW	0.0271	0.0870	0.1436	0.3861	0.5137	0.5837	0.8669	0.9193	0.9365

Table 4.11: Cluster Size and Location Estimates by 2-dim Scan Statistics, σ_0^2 Known, $m = 10, (a, b) = (11, 11)$

	$M = 100$		$M = 250$	
	$\hat{m}$	$(\hat{a}, \hat{b})$	$\hat{m}$	$(\hat{a}, \hat{b})$
MW	13.78	(9.17, 9.14)	13.68	(9.18,9.22)
VW	9.95	(11.03,11.02)	9.96	(11.05,11.05)

Table 4.12: Cluster Size and Location Estimates by 2-dim Conditional Scan Statistics, σ_0^2 Unknown, $m = 10, (a, b) = (11, 11)$

	$M = 100$		$M = 250$	
	$\hat{m}$	$(\hat{a}, \hat{b})$	$\hat{m}$	$(\hat{a}, \hat{b})$
MW	13.77	(9.17,9.15)	13.57	(9.24,9.28)
VW	9.94	(11.04,11.04)	9.94	(11.05,11.06)

4.6 Concluding Remarks

In this chapter we have investigated the performance of fixed, multiple and variable window scan statistics for detecting a local change in variance for i.i.d. normal observations in a two dimensional rectangular region. These scan statistics have been derived for the case when σ_0^2 is known and unknown, where σ_0^2 is the variance under the null hypothesis of no local change in variance. We have evaluated the performance of these scan statistics via simulation. Based on simulation results presented in Section 4.5, one can conclude that when the local shift in variance is moderate or large and the size of the rectangular window where the change in variance has occurred is unknown, the variable and multiple window scan statistics performed well. The variable window scan statistic performed somewhat better than the multiple window scan statistic when the correct window size, where the change in variance has occurred is not included in the sequence of scanned windows of the multiple window scan statistic. Hence, we recommend the use of the variable window scan statistics for moderate size M of the rectangular region. For $M \geq 500$, the computing for the variable window scan statistic will be too intense, in which case we recommend the use of the multiple window scan statistic, whose implementation will be much faster with only a small loss of power.

We would like to add that both multiple and variable window scan statistics, and their conditional versions for the unknown σ_0^2 case, can be used to estimate the location and size of the rectangular region where a local change in variance has

occurred. In our simulation study, when the shift in variance is large, the variable window scan statistics performed well. They outperformed the multiple window scan statistics. Additional research is needed to investigate the performance of variable and multiple window scan statistics for detecting accurately the location and size of the region where a local change in variance has occurred.

Chapter 5

Summary

We have investigated the scan statistics for detecting a local change in variance for normal observations in both one and two dimensional regions, and for both the case when σ_0^2 is known and unknown, where σ_0^2 is the population variance under the null hypothesis of no local change in variance. When the correct size of the window where a local change has occurred is known, the fixed window scan statistics with the correct window size are proposed. Approximations for the distributions of the fixed window scan statistics are investigated. When the correct window size is unknown, multiple window scan statistics via P_{min} approach and variable window scan statistics based on the generalized likelihood ratio tests are developed to reduce the power loss caused by using fixed window scan statistics with incorrect window sizes. When σ_0^2 is unknown, a training sample approach, a conditional approach and a parametric bootstrap testing approach are proposed to implement the fixed window scan statistics, among which the conditional approach is suggested, based on simulation results. In addition, conditional multiple and variable window scan statistics are derived, in case both the correct window size

where the local change has occurred and σ_0^2 are unknown.

For moderate or large shift in variance in both one and two dimensional cases, simulation studies suggest that, the fixed window scan statistics with correct window size performed well. When the correct window size where a local change has occurred is unknown, both the multiple and variable window scan statistics performed well. Variable window scan statistics performed slightly better than multiple window scan statistics. Hence we recommend the use of the variable window scan statistics for small or moderate size of scanning region. For large scanning regions (e.g., $M > 500$), the multiple window scan statistics are suggested, whose implementation will be much faster with only a small loss of power. In addition, both multiple and variable window scan statistics can be used to estimate the location and size of the region where a local change in variance has occurred. For large shift in variance, simulation studies showed, the variable window scan statistics performed well and outperformed the multiple window scan statistics for providing the estimation of location and size. When σ_0^2 is also unknown, the conditional version of the fixed, multiple and variable window scan statistics showed similar results. Approximations for the distributions of the fixed window scan statistics for both one and two dimensional cases showed accurate results as well.

To further enhance the results in this dissertation, future works are needed to investigate the detection for the local change in variance and estimation of the

window location and size when the local shift in variance is small, and to investigate more approximations or algorithms that can further improve the implementation speed of the multiple and variable window scan statistics for observations in large scanning regions.

Bibliography

[1] Alexandru Amărioarei and Cristian Preda. Approximations for two-dimensional discrete scan statistics in some block-factor type dependent models. *Journal of Statistical Planning and Inference*, 151:107–120, 2014.

[2] N. Balakrishnan and M.V. Koutras. *Runs and Scans with Applications*. Wiley Series in Probability and Statistics. Wiley, New york, 2011.

[3] P. Bauer and P. Hackl. The use of mosums for quality control. *Technometrics*, 20:431–436, 1978.

[4] P. Bauer and P. Hackl. An extension of the mosum technique to quality control. *Technometrics*, 22:1–7, 1980.

[5] Jie Chen and Joseph Glaz. Two-dimensional discrete scan statistics. *Statistics & Probability Letters*, 31(1):59–68, 1996.

[6] Jie Chen and Joseph Glaz. Scan statistics for monitoring data modeled by a negative binomial distribution. In *Proceedings of the XV International Symposium on Applied Stochastic Models and Data Analysis (in press)*, Barcelona, Spain, June 2013.

[7] Russell Davidson and James G. MacKinnon. Bootstrap tests: How many bootstraps? Working Papers 1036, Queen's University, Department of Economics, March 2001.

[8] A.C. Davison and D.V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[9] A.P. Dempster. *Elements of Continuous Multivariate Analysis*. Series in behavioral sciences. Addison-Wesley, Reading, MA., 1969.

[10] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

[11] J.C. Fu and W.Y.W. Lou. *Distribution Theory of Runs and Patterns and Its Applications: A Finite Markov Chain Imbedding Approach*. World Scientific, 2003.

[12] Axel Gandy and Patrick Rubin-Delanchy. An algorithm to compute the power of monte carlo tests with guaranteed precision. *Annals of Statistics*, 41(1):125–142, 02 2013.

[13] A. Genz and F. Bretz. *Computation of Multivariate Normal and T Probabilities*. Springer, 2009.

[14] J. Glaz and N. Balakrishnan, editors. *Scan Statistics and Applications*. Statistics for industry and technology. Birkhäuser, Boston, 1999.

[15] J Glaz, J Naus, and S Wallestein. *Scan Statistics*. Springer, New York, 2001.

[16] J. Glaz, V. Pozdnyakov, and S. Wallenstein. *Scan Statistics: Methods and Applications*. Statistics for Industry and Technology. Birkhäuser, Boston, 2009.

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

[18] Joseph Glaz. Discrete scan statistics with applications to minefield detection. In *Aerospace/Defense Sensing and Controls*, pages 420–429. International Society for Optics and Photonics, 1996.

[19] Joseph Glaz, Joseph Naus, and Xiao Wang. Approximations and inequalities for moving sums. *Methodology and Computing in Applied Probability*, 14(3):597–616, 2012.

[20] Marco Guerriero, Peter Willett, and Joseph Glaz. Distributed target detection in sensor networks using scan statistics. *Signal Processing, IEEE Transactions on*, 57(7):2629–2639, 2009.

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

[22] G. Haiman and C. Preda. 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, 2002.

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

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

[25] Der-Ann Hsu. Tests for variance shift at an unknown time point. *Applied Statistics*, pages 279–284, 1977.

[26] Carla Inclan and George C Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994.

[27] S. Kotz, N. Balakrishnan, and N.L. Johnson. *Continuous Multivariate Distributions, Models and Applications*. Continuous Multivariate Distributions. Wiley, New York, 2004.

[28] M. Kulldorff. Spatial scan statistics: models, calculations and applications. *Scan Statistics and Applications*, 15:303–322, 1999.

[29] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and Methods*, 26(6):1481–1496, 1997.

[30] Martin Kulldorff and Neville Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in medicine*, 14(8):799–810, 1995.

[31] N. Mukhopadhyay. *Probability and Statistical Inference*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2000.

[32] N. Nagarwalla. A scan statistic with a variable window. *Statistics in Medicine*, 15:845–850, 1996.

[33] L.F. Robinson, V.H. de La Peña, and Y. Kushnir. Detecting shifts in correlation and variability with application to enso-monsoon rainfall relationships. *Theoretical and Applied Climatology*, 94(3-4):215–224, 2008.

[34] Andreu Sans, Vicent Arag, and Josep Llus Carrion. Testing for changes in the unconditional variance of financial time series. DEA Working Papers 5, Universitat de les Illes Balears, Departament d'Economa Aplicada, November 2003.

[35] G. William Schwert. Why does stock market volatility change over time? *The Journal of Finance*, 44(5):1115–1153, 1989.

[36] Andrew R Solow. Detecting changes through time in the variance of a long-term hemispheric temperature record: An application of robust locally weighted regression. *Journal of Climate*, 1(3):290–296, 1988.

[37] Y.L. Tong. *The Multivariate Normal Distribution*. Springer Series in Statistics. Springer, New York, 2012.

[38] Loredana Ureche-Rangau and Franck Speeg. A simple method for variance shift detection at unknown time points. *Economics Bulletin*, 31(3):2204–2218, 2011.

[39] Xiao Wang and Joseph Glaz. Variable window scan statistics for normal data. *Communications in Statistics - Theory and Methods*, 43(10-12):2489–2504, 2014.

[40] Xiao Wang, Bo Zhao, and Joseph Glaz. A multiple window scan statistic for time series models. *Statistics & Probability Letters*, 94:196–203, 2014.

[41] Bo Zhao and Joseph Glaz. Scan statistics for detecting a local change in variance for normal data with known variance. Technical Report 20, Department of Statistics, University of Connecticut, 2014.

[42] Bo Zhao and Joseph Glaz. Scan statistics for detecting a local change in variance for normal data with unknown population variance. Technical Report 9, Department of Statistics, University of Connecticut, 2015.

[43] Bo Zhao and Joseph Glaz. Scan statistics for detecting a local change in variance for two dimensional normal data. Technical Report 13, Department of Statistics, University of Connecticut, 2015.