8-30-2013

Scan Statistics for Normal Data

Xiao Wang
wxiao.stat@gmail.com

Scan Statistics for Normal Data

Xiao Wang, PhD
University of Connecticut, 2013

ABSTRACT

In this dissertation we derive accurate approximations and inequalities for the distribution of fixed window scan statistics for observations from a continuous model. Employing the R algorithms for multivariate normal and t probabilities developed by Genz and Bretz (2009), these approximations and inequalities are applied to normal observations, with mean and variance being both known and unknown. These approximations are utilized to investigate the performance of fixed window scan statistics for detecting a local shift in the process mean for iid normal data. Both one and two dimensional scan statistics are investigated. To detect a local change of unknown size in the process mean, a multiple window scan statistic is introduced and compared with fixed window scan statistics via a power comparison. These results are also extended to ARMA time series data, which consists of dependant observations. It is concluded that both approximations and inequalities are quite accurate, and when the size of a local change in the process mean is unknown, the multiple window scan statistic outperforms fixed window scan statistics.

Scan Statistics for Normal Data

Xiao Wang

B.S., Applied Mathematics and Statistics, Beijing Normal University, China, 2009

M.S., Statistics, University of Connecticut, CT, USA, 2012

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

2013

APPROVAL PAGE

Doctor of Philosophy Dissertation

Scan Statistics for Normal Data

Presented by

Xiao Wang, B.S. Applied Mathematics and Statistics, M.S. Statistics

Major Advisor _____

Joseph Glaz

Associate Advisor _____

Zhiyi Chi

Associate Advisor _____

Nitis Mukhopadhyay

University of Connecticut

2013

Acknowledgements

First of all, I'd like to thank Professor Joseph Glaz for all the guidance, encouragement and support I have received from him in past four years of my life in University of Connecticut. I have learned so much from him about research, teaching, balancing work and life. Dr Glaz's contributions have gone beyond academics, and I will always remember his words for the rest of my life. I consider myself extremely fortunate to have him as my advisor, mentor and friend.

I'd also like to thank my associate advisors, Professor Zhiyi Chi and Professor Nitis Mukhopadhyay for their contributions as members of my dissertation committee. Their advice greatly enriched the content and improved the quality of the dissertation. The Department of Statistics has provided me with an excellent atmosphere for studying, teaching and research. I must thank all the professors, staff and friends in this department, because my work would not have been possible without them.

Last but not least, I'd like to thank my parents for all the love and support they have given me over the years. They have always been the light during difficult times, and encouraged me to pursue a happier life. I express deepest love for my mother and father.

Contents

Chapter 1

Introduction

Approximations for the distribution of moving sums of independent and identically distributed (iid) random variables have been of interest in probability and statistics (Bauer and Hackl, 1978; Chan, 2009; Chu et al., 1995; Glaz and Johnson, 1988; Glaz and Naus, 1991; Glaz et al., 2001; Holst and Janson, 1990; Lai, 1974; Xia et al., 2009). Applications to quality control based on moving sums of iid normal random variables have been discussed in Bauer and Hackl (1980), Westlund (1984) and Waldman (1986). Additional research related to development of tests for detection of structural changes in an observed process of recurrent residuals has been reported in Chu et al. (1995) and Xia et al. (2009). These testing procedures are closely related to a testing procedure based on scan statistics, which is defined as the maximum of moving sums for a sequence of observations when scanning with a window of fixed length.

Most of the research on scan statistics has been focused on one dimensional case and in particular on discrete models. The connections between moving sums and fixed window scan statistics in the one dimensional case have been discussed by several authors, including Glaz and Naus (1991), Glaz et al. (2001, 2012) and Haiman (2007).

In the two dimensional case this connection has been discussed, among others by Alm (1997, 1998, 1999), Chen and Glaz (1996, 1999, 2002, 2009), and Haiman and Preda (2002, 2006). In this dissertation accurate approximations and inequalities are derived for the distribution of both one and two dimensional fixed window scan statistics for iid observations of continuous random variables. These approximations and inequalities are evaluated for normal data. Based on these approximations we will also investigate the performance of one and two dimensional multiple window scan statistics. These multiple window scan statistics have mostly been investigated for discrete data in Glaz and Zhang (2004); Chen and Glaz (2009).

The importance of scan statistics and their applications have been noted in many areas of science and technology, including: astronomy (Darling and Waterman, 1986), computer science (Pfaltz, 1983), ecology (Cressie, 1991; Koen, 1991), epidemiology (Cressie, 1991; Kulldorff, 1997), image analysis (Rosenfeld, 1978), pattern recognition (Panayirci and Dubes, 1983), material science (Alm, 1999) , minefield detection via remote sensing (Glaz, 1996), signal detection in a sensor network (Guerriero et al., 2009; Song et al., 2012), quality control (Bauer and Hackl, 1978, 1980), and reliability theory (Boutsikas and Koutras, 2000; Fu and Koutras, 1994; Malinowski and Preuss, 1995; Saperstein, 1976). In this dissertation, we will address the implementation of our methodology to time series models. Time series models have been employed in many areas of science and technology, including: stock market price fluctuation in finance (Amihud, 2002), global warming in environmental science (Webster et al., 2005), fMRI imaging in medicine

(Friston et al., 2011) and disease case study in epidemiology (Dominici et al., 2002). While monitoring data generated by stationary time series models, such as autoregressive moving average (ARMA) models, the scientists might be interested in detecting a local change in the observed data, triggered by a local change in the mean of the Gaussian white noise component of the ARMA models. In this dissertation we will propose to investigate the performance of a multiple window scan statistic for monitoring time series data for an occurrence of a local change in the process mean.

This dissertation is structured as follows.

In Chapter 2, we introduce one dimensional fixed window scan statistics as the maximum of all moving sums for a sequence of observations, when scanning with a window of fixed length. We derive approximations and inequalities for the distribution, expected stopping time and variance of the stopping time associated with moving sums of independent and identically distributed continuous observations. Numerical results are presented for a normal model, with both known and unknown mean and variance.

In Chapter 3, we generalize the methodology from the one dimensional case to the two dimensional case, where a scan statistic is defined similarly for a rectangular region. A Markov-like product-type approximation is derived for two dimensional fixed window scan statistics, whose accuracy is compared with another approximation established in Haiman and Preda (2006). Second-order Bonferroni-type inequalities are also derived, which can be used to evaluate the accuracy of the approximations. The accuracy of these approximations and inequalities is investigated for a normal model, where mean and

variance being both known and unknown are discussed. Based on approximations for the distributions of one and two dimensional fixed window scan statistics, multiple window scan statistics are introduced as the minimum p-value from multiple fixed window scan statistics. We investigate the performance of these multiple window scan statistics as test statistics for detection of a local change in the mean of a normal distribution, when the size of a change is unknown. By utilizing R algorithms for the multivariate normal and t distributions established in Genz and Bretz (2009), numerical results are presented to evaluate the efficiency of implementing the multiple window scan statistics and compare their performance, via power calculations, with fixed window scan statistics.

Finally in Chapter 4 we implement our methodology for ARMA time series models. We extend the approximations based on iid observations in Chapter 2 to correlated observations from time series models. Scan statistics are investigated for ARMA models with selected parameters and a Gaussian white noise. Based on the covariance structure for the time series model, the new R algorithms for the multivariate normal distributions established in Genz and Bretz (2009) provide readily available numerical results for the approximations. The accuracy of these approximations is verified via a simulation study. The multiple window scan statistic from Chapter 3 is also implemented for the detection of a change of unknown size in the mean of Gaussian white noise. Numerical results are presented to evaluate the efficiency of implementing the multiple window scan statistic and compare its performance, via power calculations, with fixed window scan statistics. We also illustrate the use of the multiple window scan statistic for a data set in Box and

Jenkins (1976).

Chapter 2

One Dimensional Scan Statistics

2.1 Introduction

Let $X_1, ..., X_M, ...$ be a sequence of iid normal observations with mean μ and variance σ^2. Let $Y_{r,u} = \sum_{i=r}^{u} X_i$ for $u \geq r \geq 1$. For integers $2 \leq m < M$, where m is the length of the sliding window and M is the specified range of the monitored process, define the *scan statistic*

$$S_{m,M} = \max_{m \leq j \leq M} \{Y_{j-m+1,j}\}. \tag{2.1}$$

The sequence $\{Y_{j-m+1,j}; m \leq j \leq M\}$, based on which the scan statistic is defined, contains $M - m + 1$ moving sums of length m. The random variables $\{Y_{j-m+1,j}; m \leq j \leq M\}$ have a joint multivariate normal distribution with mean vector $(m\mu, ..., m\mu)'$ and variance and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, where

$$\sigma_{i,j} = \begin{cases} (m - |i-j|)\sigma^2 & \text{when} \quad |i-j| < m \\ 0 & \text{when} \quad |i-j| \geq m \end{cases}$$

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

$$G_{m,t}(M) = P\left(Y_{1,m} < t, Y_{2,m+1} < t, ..., Y_{M-m+1,M} < t\right). \tag{2.2}$$

The distribution of the scan statistic $S_{m,M}$ is given by

$$P\left(S_{m,M} < t\right) = G_{m,t}(M). \tag{2.3}$$

Equivalently, the probability that the scan statistic exceeds level t is given by

$$P\left(S_{m,M} \geq t\right) = 1 - G_{m,t}(M). \tag{2.4}$$

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. This scan statistic can be used in detecting a local change in the process mean within a sequence of M observations via testing the null hypothesis of randomness, H_0, that assumes $X_i, 1 \leq i \leq M$, are iid normal random variables with mean μ_0 and variance σ^2. For the alternative hypothesis, H_1, of a local change in μ , one often specifies a segment of m consecutive observations $R(i_0, m) = \{i_0, i_0 + 1, ..., i_0 + m - 1\}$, where $1 \leq i_0 \leq M - m + 1$ is unknown and $2 \leq m \leq M/4$ is a specified window length. Under H_1, for any $i_0 \leq i \leq i_0 + m - 1$, X_i has a normal distribution with mean μ_1 and variance σ^2, where $\mu_1 > \mu_0$. For $i \notin R(i_0, m)$, $X_i's$ are distributed according to the distribution specified by the null hypothesis. When the length of the sliding window m

is known, the generalized likelihood ratio test rejects the null hypothesis of randomness in favor of the local change alternative hypothesis H_1, whenever $S_{m,M}$ exceeds the value t, where t is determined from $P(S_{m,M} \geq t|H_0) = \alpha$, where α is a specified significance level of the testing procedure (Glaz et al., 2001).

Approximations for the distribution of moving sums of iid discrete random variables have been investigated in Glaz and Naus (1991). Accurate approximations, via extremes of a 1-dependent stationary sequence, have been derived in Haiman (2007). In this chapter, we extend the approach in Glaz and Naus (1991) to moving sums of iid continuous random variables. In Section 2.2, Theorem 1, we present inequalities for the distribution of moving sums of iid random variables. In Section 2.3, we present approximations for $G(M)$. In Sections 2.4 and 2.5, approximations and inequalities are derived for the expected time and variance of the time it takes a moving sum of length m to cross a specified level t, respectively. In Section 2.6, we describe how to implement the results presented in Sections 2.1-2.5 for a sequence of iid normal random variables with mean μ and variance σ^2. Both cases of μ and σ^2 being known and unknown are discussed. For the case of σ^2 unknown, Theorem 2 confirms that the approximations and inequalities derived in Sections 2.1-2.5 remain valid. In turn, these approximations can be employed to implement the scan statistic defined in (2.1).

The new R algorithms for multivariate normal and t distributions (Genz and Bretz, 2009) provide readily available numerical results to evaluate upper tail probabilities for the scan statistic. In Section 2.7, we present numerical results for a scan statistic based

on the inequalities and approximations discussed in this article. Concluding remarks are given in Section 2.8.

2.2 Inequalities for $G(M)$

Let $X_1, ..., X_M$ be iid continuous random variables with mean μ and variance σ^2. Let $Y_{r,u} = \sum_{i=r}^{u} X_i$ for $u \geq r \geq 1$, and $Y_{r,u} = -\infty$, otherwise. For integers $2 \leq m \leq R \leq U \leq M$ and $-\infty < t < \infty$, define

$$N_{R,U} = \max_{R \leq j \leq U} \{Y_{j-m+1,j}\},$$

to be the largest element in a sequence of moving sums of length m. Let

$$\tau_{m,t} = \inf \{k \geq m; Y_{k-m+1,k} \geq t\},$$

be the waiting time time for a sequence of moving sums of length m to exceed a level t. For $0 \leq j \leq m-1$, define $G(j) = 1$. For $M \geq m$, let

$$G_{m,t}(M) = P(\tau_{m,t} > M) = P(N_{m,M} < t),$$

denote the probability of no exceedance of level t within a block of $M-m+1$ consecutive moving sums of length m. Let

$$f_{m,t}(k) = P\left(\tau_{m,t} = k\right),$$

be the probability that the first exceedance of level t by a sequence of moving sums of length m occurs at time k. When the values of m and t are clearly understood we will abbreviate $\tau_{m,t}$, $G_{m,t}(M)$ and $f_{m,t}(k)$ to τ, $G(M)$ and $f(k)$, respectively. The following two lemmas are needed to establish probability inequalities for $G(M)$.

Lemma 1 *For integers $L \geq 1$ and $k \geq Lm$,*

$$f(k) \leq G(k - Lm)f(Lm). \tag{2.5}$$

Proof For $k \geq (L+1)m$,

$$\begin{aligned}
f(k) &= P\left\{(N_{m,k-1} < t) \cap (Y_{k-m+1,k} \geq t)\right\} \\
&= P\left\{(N_{m,k-Lm} < t) \cap (N_{k-Lm+1,k-1} < t) \cap (Y_{k-m+1,k} \geq t)\right\} \\
&\leq P\left\{(N_{m,k-Lm} < t) \cap \left(N_{k-(L-1)m,k-1} < t\right) \cap (Y_{k-m+1,k} \geq t)\right\} \\
&= P\left(N_{m,k-Lm} < t\right) P\left\{\left(N_{k-(L-1)m,k-1} < t\right) \cap (Y_{k-m+1,k} \geq t)\right\} \\
&= G(k - Lm)f(Lm).
\end{aligned}$$

where the last step follows from the stationary property of the moving sums of iid random variables.

To verify inequality (2.5) for $Lm \leq k \leq (L+1)m-1$, note that $G(j) = 1$ for $0 \leq j \leq m-1$. Therefore, for $L = 1$ and $2 \leq m \leq k \leq 2m-1$,

$$\begin{aligned} f(k) &= P\left\{(N_{m,k-1} < t) \cap (Y_{k-m+1,k} \geq t)\right\} \\ &\leq P\left(Y_{k-m+1,k} \geq t\right) = P\left(Y_{1,m} \geq t\right) = f(m) \\ &= G(k-m)f(m). \end{aligned}$$

For $L \geq 2$,

$$\begin{aligned} f(k) &= P\left\{(N_{m,k-1} < t) \cap (Y_{k-m+1,k} \geq t)\right\} \\ &\leq P\left\{(N_{k-(L-1)m,k-1} < t) \cap (Y_{k-m+1,k} \geq t)\right\} \\ &= f(Lm) = G(k-Lm)f(Lm). \end{aligned}$$

This concludes the proof of Lemma 1.

Lemma 2 *For integers $L \geq 1$ and $k \geq (L+1)m$,*

$$f(k) \geq G(k-Lm)f((L+1)m). \tag{2.6}$$

Proof Let

$$E_1 = (N_{m,k-Lm} < t)$$

and

$$E_2 = \{(N_{k-Lm,k-1} < t) \cap (Y_{k-m+1,k} \geq t)\}.$$

Let

$$S = (X_{k-Lm+1}, ..., X_k).$$

Conditioned on S, the indicator random variables I_{E_1} and I_{E_2} are decreasing functions of $X_{1,}....,X_{k-Lm}$. Therefore, if follows from J.D. Esary and Walkup (1967), Section 2, that

$$P(E_1 \cap E_2|S) \geq P(E_1|S)P(E_2|S) = P(E_1)P(E_2|S).$$

By Averaging over the density of S we get

$$P(E_1 \cap E_2) \geq P(E_1)P(E_2) = G(k - Lm)f((L + 1)m).$$

This concludes the proof of Lemma 2.

Now we state a theorem about lower and upper bounds for $G(M)$ for iid continuous random variables. The case of iid integer valued random variables has been investigated in Glaz and Naus (1991).

Theorem 1 *For integers* $i, m \geq 2$ *and* $L \geq 1$:

$$G(M) \geq \frac{G(im)}{\left[1 + \frac{G(Lm-1)-G(Lm)}{G((L+1)m-1)}\right]^{M-im}}, \qquad \text{for} M \geq (i \vee L)m, \tag{2.7}$$

and

$$G(M) \leq G(im)\{1-[G((L+1)m-1)-G((L+1)m)]\}^{M-im}, \qquad \text{for } M \geq (i \vee (L+1))m. \tag{2.8}$$

Proof We first derive a lower bound for $G(M)$. For $k \geq Lm$, it follows from Lemma 1 that

$$f(k) = G(k-1) - G(k) \leq G(k-Lm)f(Lm).$$

The indicator random variables of the events $(N_{m,u} < t)$ and $(N_{u-m+2,u+v} < t)$ are decreasing functions of $X_i's$. It follows J.D. Esary and Walkup (1967), Section 2, that for $u \geq m$ and $v \geq 0$

$$\begin{aligned} G(u+v) &= P\left\{(N_{m,u} < t) \cap (N_{u-m+2,u+v} < t)\right\} \\ &\geq P(N_{m,u} < t)P(N_{u-m+2,u+v} < t) \\ &= G(u)G(v+m-1). \end{aligned}$$

In particular, for $k \geq Lm, u = k - Lm$ and $v = Lm$, the inequality above implies that

$$G(k - Lm) \leq G(k)/G((L + 1)m - 1).$$

Therefore,

$$G(k - 1) - G(k) \leq f(Lm)G(k)/G((L + 1)m - 1),$$

leading to

$$G(k - 1) \leq (1 + B_{1,L})G(k),$$

or equivalently

$$G(k - 1)(1 + B_{1,L})^{k-1} \leq G(k)(1 + B_{1,L})^{k},$$

where

$$B_{1,L} = \frac{f(Lm)}{G((L + 1)m - 1)}.$$

The last inequality implies that

$$G(k)(1 + B_{1,L})^{k}$$

is an increasing function in k. Therefore, for $L \geq 1$ and $M \geq (i \vee L)m$,

$$G(M)(1 + B_{1,L})^{M} \geq G(im)(1 + B_{1,L})^{im}$$

or equivalently

$$\begin{aligned} G(M) &\geq G(im)/(1+B_{1,L})^{M-im} \\ &= \frac{G(im)}{\left[1+\frac{f(Lm)}{G((L+1)m-1)}\right]^{M-im}} \\ &= \frac{G(im)}{\left[1+\frac{G(Lm-1)-G(Lm)}{G((L+1)m-1)}\right]^{M-im}}. \end{aligned}$$

To derive the upper bound for $G(M)$, note that Lemma 2 implies that for $k \geq (L+1)m$,

$$f(k) = G(k-1) - G(k) \geq f((L+1)m)G(k-Lm) \geq f((L+1)m)G(k-1). \quad (2.9)$$

For $L \geq 1$, let

$$A_{1,L} = f(L+1)m).$$

It follows from inequality (2.9) that

$$G(k) \leq G(k-1)(1-A_{1,L})$$

or equivalently

$$G(k)/(1-A_{1,L})^k \leq G(k-1)/(1-A_{1,L})^{k-1}.$$

Therefore,

$$G(k)/(1-A_{1,L})^k$$

is a decreasing function of k. This fact implies that for $L \geq 1$ and $M \geq (i \vee (L+1))m$,

$$G(M)/(1-A_{1,L})^M \leq G(im)/(1-A_{1,1})^{im}$$

or equivalently

$$\begin{aligned} G(M) &\leq G(im)[1-f((L+1)m)]^{M-im} \\ &= G(im)\{1-[G((L+1)m-1)-G((L+1)m)]\}^{M-im}. \end{aligned}$$

This concludes the proof of Theorem 1.

Remark. The proofs of Lemma 1, Lemma 2 and Theorem 1 follow closely the approach in Glaz and Naus (1991), where similar probability inequalities were derived for moving sums of iid integer valued random variables.

For computing the bounds for $G(M)$, we employ, $i = 2$ and $L = 2$ for the lower bound and $i = 2$ and $L = 1$ for the upper bound. For these choices of i and L, and since $G(3m-1) \geq G(2m-1)G(2m)$, the inequalities in Theorem 1 reduce to:

$$G(M) \geq \frac{G(2m)}{\left[1+\frac{G(2m-1)-G(2m)}{G(2m-1)G(2m)}\right]^{M-2m}}, \qquad M \geq 2m, \tag{2.10}$$

and

$$G(M) \leq G(2m)\{1-[G(2m-1)-G(2m)]\}^{M-2m}, \qquad M \geq 2m. \tag{2.11}$$

To prove that the upper bound is always larger than the lower bound, one needs to show the following inequality

$$1 - [G(2m-1) - G(2m)] \geq \frac{1}{1 + \frac{G(2m-1)-G(2m)}{G(2m-1)G(2m)}}$$
$$= \frac{G(2m-1)G(2m)}{G(2m-1)G(2m) + G(2m-1) - G(2m)}.$$

Multiply both sides of the inequality by $G(2m-1)G(2m) + G(2m-1) - G(2m)$ and factor out $G(2m-1)G(2m)$, one gets

$$[G(2m-1) - G(2m)][1 - G(2m-1)G(2m) - G(2m-1) + G(2m)] \geq 0$$

which is equivalent to

$$[1 - G(2m-1)][1 + G(2m)] \geq 0$$

which holds always and we expect these bounds to be tight for a large value of t, since they converge as $G(2m) \to 1$ and $G(2m-1) - G(2m) \to 0$, which is satisfied as $t \longrightarrow \infty$.

2.3 Approximations for $G(M)$

We now proceed to derive a Markov-type approximation for $G(M)$ following a method introduced in Naus (1982). Let $M = Km + \upsilon$, where $K \geq 3, m \geq 2$ and $0 \leq \upsilon \leq m-1$

are integers. Then, for $2 \leq L \leq K-1$

$$\begin{aligned} G(M) &= P\left\{\max_{m\leq k\leq M} Y_{k-m+1,k} < t\right\} = P\left(\bigcap_{j=1}^{K} E_j\right) \\ &= P\left(\bigcap_{i=1}^{L-1} E_i\right) \prod_{j=L}^{K} P\left(E_j | \bigcap_{h=1}^{j-1} E_h\right), \end{aligned} \tag{2.12}$$

where for $1 \leq j \leq K-1$

$$E_j = \left(\max_{jm\leq k\leq (j+1)m} Y_{k-m+1,k} < t\right),$$

which can be interpreted as the event of no exceedance of level t within a block of $m+1$ consecutive partial sums of length m, and

$$E_K = \left(\max_{Km\leq k\leq Km+\upsilon} Y_{k-m+1,k} < t\right).$$

By conditioning on most recent past of $L \geq 2$ events E_j in (2.12), we propose the following approximation for $G(M)$:

$$\begin{aligned} G(M) &\approx P\left(\bigcap_{i=1}^{L-1} E_i\right)\left[\prod_{j=L}^{K-1} P\left(E_j \Big| \bigcap_{h=j-L+1}^{j-1} E_h\right)\right] P\left(E_K \Big| \bigcap_{p=K-L+1}^{K-1} E_p\right) \\ &= P\left(\bigcap_{i=1}^{L} E_i\right)\left\{\prod_{j=L+1}^{K-1}\left[\frac{P\left(\bigcap_{h=j-L+1}^{j} E_h\right)}{P\left(\bigcap_{h=j-L+1}^{j-1} E_h\right)}\right]\right\}\frac{P\left(\bigcap_{p=K-L-1}^{K} E_p\right)}{P\left(\bigcap_{p=K-L1}^{K-1} E_p\right)} \\ &= G((L+1)m)\left[\frac{G((L+1)m)}{G(Lm)}\right]^{K-L-1}\frac{G(Lm+\upsilon)}{G(Lm)}. \end{aligned} \tag{2.13}$$

Equation (2.13) follows from the fact that for $2 \leq L \leq K-1$, the events $\{E_j; 1 \leq j \leq K-1\}$ and $\{E_{j-L+1} \cap \cdots \cap E_j; 2 \leq j \leq K-L+1\}$ are stationary and that for $1 \leq L \leq K-2$, $P\left(\bigcap_{i=1}^{L} E_i\right) = G((L+1)m)$ and $P(E_{K-L} \cap \cdots \cap E_K) = G((L+1)m+\upsilon)$. For $L = 2$, the above approximation reduces to:

$$G(M) \approx G(3m)\left[\frac{G(3m)}{G(2m)}\right]^{K-3}\frac{G(2m+\upsilon)}{G(2m)}.$$

If $M = Km$, approximation (2.13) simplifies to

$$G(M) \approx G((L+1)m)\left[\frac{G((L+1)m)}{G(Lm)}\right]^{K-L-1}. \tag{2.14}$$

For $M = Km$ and $L = 2$ this approximation reduces to:

$$G(M) \approx G(3m) \left[\frac{G(3m)}{G(2m)}\right]^{K-3}. \tag{2.15}$$

Haiman (1999, 2007) derived accurate approximations for $G(M)$ for iid discrete random variables. These approximations are valid as well for iid continuous random variables. A nice feature of these approximations is that a sharp error bound can be easily evaluated. For the problem at hand, for any t and $M \geq 3m$, such that $1-G(2m) \leq .025$ and $3.3M[1 - G(2M)]^2 \leq 1$, the following approximation for $G(M)$ is obtained from Haiman (2007), Corollary 2 and Equation 2.2:

$$G(M) \approx \frac{2G(2m) - G(3m)}{\left[1 + G(2m) - G(3m) + 2\left(G(2m) - G(3m)\right)^2\right]^{M/m-1}}, \tag{2.16}$$

with an error bound of approximately

$$3.3[1 - G(2m)]^2(M/m - 1). \tag{2.17}$$

In Section 2.7, based on approximations (2.15) and (2.16), we evaluate approximations for the distribution of the scan statistic for a sequence of iid normal and t random variables.

2.4 Approximations and Inequalities for $E(\tau)$

Recall from Section 2.2, that τ is the waiting time for moving sums to exceed level t. Since,

$$\begin{aligned} E(\tau) &= \sum_{M=0}^{\infty} P(\tau > M) = m + \sum_{M=m}^{\infty} P(\tau > M) \\ &= m + \sum_{M=m}^{(L+1)m} G(M) + \sum_{M=(L+1)m+1}^{\infty} G(M), \end{aligned} \tag{2.18}$$

approximations and inequalities for $G(M)$ yield approximations and inequalities for $E(\tau)$.

For $L \geq 1$, it follows from (2.18) and (2.13) that an approximation for $E(\tau)$ is given by:

$$\begin{aligned} E(\tau) &\approx m + \sum_{M=m}^{(L+1)m} G(M) + \sum_{\upsilon=1}^{m} \sum_{K=L+1}^{\infty} G((L+1)m) \left[\frac{G((L+1)m)}{G(Lm)}\right]^{K-L-1} \\ &\quad \times \frac{G(Lm+\upsilon)}{G(Lm)} \\ &= m + \sum_{M=m}^{(L+1)m} G(M) + G((L+1)m) \sum_{\upsilon=1}^{m} \frac{G(Lm+\upsilon)}{G(Lm)} \sum_{K=L+1}^{\infty} \left[\frac{G((L+1)m)}{G(Lm)}\right]^{K-L-1}. \end{aligned}$$

Therefore,

$$E(\tau) \approx m + \sum_{M=m}^{(L+1)m} G(M) + G((L+1)m)\frac{G(Lm)}{G(Lm) - G((L+1)m)}$$
$$\times \sum_{\upsilon=1}^{m} \frac{G(Lm+\upsilon)}{G(Lm)}$$
$$= m + \sum_{M=m}^{Lm} G(M) + \frac{G(Lm)}{G(Lm) - G((L+1)m)} \sum_{\upsilon=1}^{m} G(Lm+\upsilon). \tag{2.19}$$

For $L = 2$, we get the following approximation for $E(\tau)$:

$$E(\tau) \approx m + \sum_{M=m}^{2m} G(M) + \frac{G(2m)}{G(2m) - G(3m)} \sum_{\upsilon=1}^{m} G(2m+\upsilon). \tag{2.20}$$

By summing the geometric series in (2.7) for $i = L \geq 2$, we get the following lower bound for $E(\tau)$:

$$E(\tau) \geq m + \sum_{M=m}^{Lm-1} G(M) + G(Lm)\left[1 + \frac{G((L+1)m-1)}{G(Lm-1) - G(Lm)}\right]. \tag{2.21}$$

By summing the geometric series in (2.8) for $i = L+1$, $L \geq 1$, we get the following upper bound for $E(\tau)$:

$$E(\tau) \leq m + \sum_{M=m}^{(L+1)m-1} G(M) + \frac{G((L+1)m)}{G((L+1)m-1) - G((L+1)m)}. \tag{2.22}$$

With $L = 2$ for the lower bound and $L = 1$ for the upper bound:

$$E(\tau) \geq m + \sum_{M=m}^{2m-1} G(M) + G(2m) \left[1 + \frac{G(2m-1)G(2m)}{G(2m-1) - G(2m)} \right] \tag{2.23}$$

and

$$E(\tau) \leq m + \sum_{M=m}^{2m-1} G(M) + \frac{G(2m)}{G(2m-1) - G(2m)}. \tag{2.24}$$

For large value of t, $G(2m-1), G(2m) \to 1$ and $G(2m-1) - G(2m) \to 0$. Therefore, the value of both bounds for $E(\tau)$ will be large and dominated by $m + \sum_{M=m}^{2m-1} G(M) + 1/[G(2m-1) - G(2m)]$. The difference between these bounds will be at most 1.

2.5 Approximations and Inequalities for $Var(\tau)$

Approximations and inequalities for $E(\tau)$ have been derived in Section 2.4. Since,

$$Var(\tau) = E[\tau(\tau - 1)] + E(\tau) - [E(\tau)]^2, \tag{2.25}$$

approximations and inequalities for $E[\tau(\tau-1)]$ will yield approximations and inequalities for $Var(\tau)$. For $L \geq 1$, the following representation for $E[\tau(\tau - 1)]$ is employed:

$$E[\tau(\tau-1)] = 2\sum_{M=1}^{\infty} MG(M) = 2\sum_{M=1}^{m-1} M + 2\sum_{M=m}^{(L+1)m} MG(M)$$
$$+ 2\sum_{M=(L+1)m+1}^{\infty} MG(M)$$
$$= m(m-1) + 2\sum_{M=m}^{(L+1)m} MG(M) + 2\sum_{M=(L+1)m+1}^{\infty} MG(M). \qquad (2.26)$$

For clarity, we defer the technical details related to the derivation of the approximation and inequalities for $E[\tau(\tau-1)]$, via the approximation and inequalities for $\sum_{M=(L+1)m+1}^{\infty} MG(M)$, to the end of this section. We present below a simple approximation and inequalities for $Var(\tau)$. This approximation is evaluated in Section 2.7.

For $L = 2$, it follows from (2.35) that

$$E[\tau(\tau-1)] \approx m(m-1) + 2\sum_{M=m}^{3m} MG(M) + \frac{2mx(3-2x)}{(1-x)^2}\sum_{\upsilon=1}^{m} G(2m+\upsilon)$$
$$+ \frac{2x}{1-x}\sum_{\upsilon=1}^{m} \upsilon G(2m+\upsilon) = \widehat{E}_{L=2}[\tau(\tau-1)],$$

where

$$x = \frac{G(3m)}{G(2m)}.$$

Therefore, for $L = 2$, an approximation for $Var(\tau)$ is given by:

$$\widehat{E}_{L=2}[\tau(\tau-1)] + \widehat{E}_{L=2}(\tau) - \left[\widehat{E}_{L=2}(\tau)\right]^2. \qquad (2.27)$$

For $L = 2$, we get from (2.40) that:

$$E[\tau(\tau-1)] \geq m(m-1) + 2\sum_{M=m}^{2m} MG(M) + \frac{2mG(2m)u(1-u^m)(3-2u^m)}{(1-u)(1-u^m)^2} + \frac{2G(2m)u[1-(m+1)u^m+mu^{m+1}]}{(1-u)^2(1-u^m)} = LB_{L=2}\left\{E[\tau(\tau-1)]\right\}, \quad (2.28)$$

where

$$u = \frac{G(2m-1)G(2m)}{G(2m-1)G(2m) + G(2m-1) - G(2m)}.$$

Hence:

$$Var(\tau) \geq LB_{L=2}\left\{E[\tau(\tau-1)]\right\} + LB_{L=2}E(\tau) - \left[UB_{L=1}E(\tau)\right]^2. \quad (2.29)$$

For $L = 1$, inequality (2.38) implies:

$$E[\tau(\tau-1)] \leq m(m-1) + 2\sum_{M=m}^{2m} MG(M) + \frac{2mG(2m)y(1-y^m)(2-y^m)}{(1-y)(1-y^m)^2} + \frac{2G(2m)y[1-(m+1)y^m+my^{m+1}]}{(1-y)^2(1-y^m)} = UB_{L=1}\left\{E[\tau(\tau-1)]\right\}, \quad (2.30)$$

where

$$y = 1 - [G(2m-1) - G(2m)].$$

This yields the following upper bound:

$$Var(\tau) \leq UB_{L=1}\left\{E[\tau(\tau-1)]\right\} + UB_{L=1}E(\tau) - \left[LB_{L=2}E(\tau)\right]^2. \quad (2.31)$$

We now derive approximations and inequalities for $\sum_{M=(L+1)m+1}^{\infty} MG(M)$ and $E[\tau(\tau-1)]$, based on which approximations and inequalities for $Var(\tau)$ have been obtained.

To approximate $\sum_{M=(L+1)m+1}^{\infty} MG(M)$, let $M = Km+\upsilon$, where $K \geq 2$ and $1 \leq \upsilon \leq m$. Then,

$$\sum_{M=(L+1)m+1}^{\infty} MG(M) \approx \sum_{\upsilon=1}^{m} \sum_{K=L+1}^{\infty} (Km+\upsilon) G((L+1)m) \times \left[\frac{G((L+1)m)}{G(Lm)}\right]^{K-L-1} \frac{G(Lm+\upsilon)}{G(Lm)}.$$

It follows that,

$$\sum_{M=(L+1)m+1}^{\infty} MG(M) \approx \frac{G((L+1)m)}{G(Lm)} \sum_{\upsilon=1}^{m} G(Lm+\upsilon) \times \left\{ m \sum_{K=L+1}^{\infty} K \left[\frac{G((L+1)m)}{G(Lm)}\right]^{K-L-1} + \upsilon \sum_{K=L+1}^{\infty} \left[\frac{G((L+1)m)}{G(Lm)}\right]^{K-L-1} \right\}.$$

To simplify the presentation of the results we will use the following notation:

$$x = \frac{G((L+1)m)}{G(Lm)}.$$

Then,

$$\sum_{M=(L+1)m+1}^{\infty} MG(M) \approx x \sum_{\upsilon=1}^{m} G(Lm+\upsilon) \left(m \sum_{K=L+1}^{\infty} K x^{K-L-1} + \upsilon \sum_{K=L+1}^{\infty} x^{K-L-1} \right).$$

Since $0 < x < 1$, the following geometric series sums to:

$$\sum_{K=L+1}^{\infty} x^{K-L-1} = \frac{1}{1-x}.$$

To sum

$$\sum_{K=L+1}^{\infty} Kx^{K-L-1},$$

we employ the following well know identity

$$\sum_{j=1}^{\infty} jx^{j-1} = \frac{1}{(1-x)^2}.$$

It follows that

$$\begin{aligned}\sum_{K=L+1}^{\infty} Kx^{K-L-1} &= \frac{1}{x^L}\sum_{K=L+1}^{\infty} Kx^{K-1} = \frac{1}{x^L}\left[\sum_{K=1}^{\infty} Kx^{K-1} - \sum_{K=1}^{L} Kx^{K-1}\right] \\ &= \frac{1}{x^L}\left[\frac{1}{(1-x)^2} - \sum_{K=1}^{L} Kx^{K-1}\right] = \frac{1}{x^L(1-x)^2} - \frac{1}{x^L}\sum_{K=1}^{L} Kx^{K-1}.\end{aligned}$$

Since

$$\begin{aligned}\sum_{K=1}^{L} Kx^{K-1} &= \frac{d}{dx}\left[\sum_{K=0}^{L} x^K\right] = \frac{d}{dx}\left[\frac{1-x^{L+1}}{1-x}\right] \\ &= \frac{1 + Lx^{L+1} - (L+1)x^L}{(1-x)^2},\end{aligned} \tag{2.32}$$

we get

$$\begin{aligned}\sum_{K=L+1}^{\infty} K x^{K-L-1} &= \frac{1}{x^L(1-x)^2}\left[(L+1)x^L - Lx^{L+1}\right] \\ &= \frac{L+1-Lx}{(1-x)^2}. \end{aligned} \tag{2.33}$$

Hence,

$$\begin{aligned}\sum_{M=(L+1)m+1}^{\infty} MG(M) &\approx x\sum_{\upsilon=1}^{m} G(Lm+\upsilon)\left[\frac{m(L+1-Lx)}{(1-x)^2} + \frac{\upsilon}{1-x}\right] \\ &= \frac{mx(L+1-Lx)}{(1-x)^2}\sum_{\upsilon=1}^{m} G(Lm+\upsilon) + \frac{x}{1-x}\sum_{\upsilon=1}^{m} \upsilon G(Lm+\upsilon)\end{aligned} \tag{2.34}$$

and

$$\begin{aligned}E[\tau(\tau-1)] \approx\ & m(m-1) + 2\sum_{M=m}^{(L+1)m} MG(M) + \frac{2mx(L+1-Lx)}{(1-x)^2}\sum_{\upsilon=1}^{m} G(Lm+\upsilon) \\ &+ \frac{2x}{1-x}\sum_{\upsilon=1}^{m} \upsilon G(Lm+\upsilon).\end{aligned} \tag{2.35}$$

We now derive bounds for $\sum_{M=(L+1)m+1}^{\infty} MG(M)$ and $E[\tau(\tau-1)]$. Let $M = Km+\upsilon$,

where $K \geq 2$ and $1 \leq \upsilon \leq m$. Note that:

$$\begin{aligned}\sum_{M=(L+1)m+1}^{\infty} MG(M) \leq & \sum_{\upsilon=1}^{m}\sum_{K=L+1}^{\infty}(Km+\upsilon)G((L+1)m) \\ & \times \{1-[G((L+1)m-1)-G(L+1)m)]\}^{Km+\upsilon-(L+1)m} \\ & = G((L+1)m)\sum_{\upsilon=1}^{m}\{1-[G((L+1)m-1)-G((L+1)m)]\}^{\upsilon} \\ & \times \sum_{K=L+1}^{\infty}(Km+\upsilon)\{1-[G((L+1)m-1)-G((L+1)m)]\}^{Km-(L+1)m}.\end{aligned} \tag{2.36}$$

Therefore,

$$\begin{aligned}\sum_{M=(L+1)m+1}^{\infty} MG(M) \leq & G((L+1)m)\sum_{\upsilon=1}^{m}\{1-[G((L+1)m-1)-G((L+1)m)]\}^{\upsilon} \\ & \times m\sum_{K=L+1}^{\infty}K\{1-[G((L+1)m-1)-G((L+1)m)]\}^{Km-(L+1)m} \\ & + G((L+1)m)\sum_{\upsilon=1}^{m}\upsilon\{1-[G((L+1)m-1)-G((L+1)m)]\}^{\upsilon} \\ & \times \sum_{K=L+1}^{\infty}\{1-[G((L+1)m-1)-G((L+1)m)]\}^{Km-(L+1)m}.\end{aligned}$$

To simplify the presentation of the results, let

$$y = 1-[G((L+1)m-1)-G((L+1)m)]$$

and

$$z = y^m.$$

Then,

$$\sum_{M=(L+1)m+1}^{\infty} MG(M) \leq mG((L+1)m)\sum_{\upsilon=1}^{m} y^{\upsilon}\sum_{K=L+1}^{\infty} Kz^{K-L-1} + G((L+1)m)\sum_{\upsilon=1}^{m} \upsilon y^{\upsilon}\sum_{K=L+1}^{\infty} z^{K-L-1}. \quad (2.37)$$

The first and fourth sums are geometric series in y and z, respectively, and have the following simple expressions:

$$\sum_{\upsilon=1}^{m} y^{\upsilon} = \frac{y(1-y^m)}{1-y}$$

and

$$\sum_{K=L+1}^{\infty} z^{K-L-1} = \frac{1}{1-z} = \frac{1}{1-y^m}.$$

It follows from identity in (2.32) that:

$$\sum_{\upsilon=1}^{m} \upsilon y^{\upsilon} = \frac{y[1-(m+1)y^m + my^{m+1}]}{(1-y)^2}.$$

Equation (2.33) implies:

$$\sum_{K=L+1}^{\infty} Kz^{K-L-1} = \frac{L+1-Lz}{(1-z)^2} = \frac{L+1-Ly^m}{(1-y^m)^2}.$$

Therefore,

$$E[\tau(\tau-1)] \leq m(m-1) + 2\sum_{M=m}^{(L+1)m} MG(M) + \frac{2mG((L+1)m)y(1-y^m)(L+1-Ly^m)}{(1-y)(1-y^m)^2} + \frac{2G((L+1)m)y[1-(m+1)y^m+my^{m+1}]}{(1-y)^2(1-y^m)}. \tag{2.38}$$

We now derive a lower bound for $\sum_{M=Lm+1}^{\infty} MG(M)$ and $E[\tau(\tau-1)]$. For $i = L$, it follows from the lower bound for $G(M)$ in (2.7) that:

$$\begin{aligned} \sum_{M=Lm+1}^{\infty} MG(M) &\geq \sum_{\upsilon=1}^{m}\sum_{K=L}^{\infty}(Km+\upsilon)\frac{G(Lm)}{\left[1+\frac{G(Lm-1)-G(Lm)}{G((L+1)m-1)}\right]^{Km+\upsilon-Lm}} \\ &\geq \sum_{\upsilon=1}^{m}\sum_{K=L}^{\infty}(Km+\upsilon)\frac{G(Lm)}{\left[1+\frac{G(Lm-1)-G(Lm)}{G(Lm-1)G(Lm)}\right]^{Km+\upsilon-Lm}} \end{aligned} \tag{2.39}$$

The right hand side of the inequality (2.39) has the same structure as the inequality (2.36). Let

$$u = \frac{G(Lm-1)G(Lm)}{G(Lm-1)G(Lm)+G(Lm-1)-G(Lm)}$$

and

$$w = u^m.$$

Therefore, for $L \geq 1$:

$$E[\tau(\tau-1)] \geq m(m-1) + 2\sum_{M=m}^{Lm} MG(M) + \frac{2mG(Lm)u(1-u^m)(L+1-Lu^m)}{(1-u)(1-u^m)^2} + \frac{2G(Lm)u[1-(m+1)u^m+mu^{m+1}]}{(1-u)^2(1-u^m)}. \quad (2.40)$$

2.6 Moving Sums for Normal Observations

Let $X_1, ..., X_M, ...$ be iid normal random variables with mean μ and variance σ^2. Without loss of generality we can assume that $\mu = 0$. Otherwise, we will consider the sequence of recurrent residuals (Bauer and Hackl (1980), Section 1.2):

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

which are iid normal random variables with mean equal to 0 and variance equal to σ^2. First, we consider the case of σ^2 being known, and without loss of generality assume $\sigma^2 = 1$. We employ a randomized quasi Monte-Carlo R algorithm for evaluating multivariate normal probabilities (Genz and Bretz, 2009) to evaluate the approximations and bounds for the distribution and moments of moving sums that have been derived in Sections 2.1-2.5. Approximations and bounds for probabilities of moving sums yield approximations and bounds for tail probabilities of the scan statistic in Equation (2.4). Numerical results for selected values of the parameters are presented in Section 2.7.

We now consider the case of σ^2 being unknown. In what follows, we will show how one can generalize Theorem 1, Section 2.2, to derive approximations and bounds for the distribution and the moments of moving sums.

Let $X_1, X_2, ...$be iid normal random variables with mean 0 and unknown variance σ^2. Assume that a training sample of n_0 iid normal random variables, independent of the X_i's, and with the same distribution as X_1, has been observed. Let S^2 be the sample variance based on n_0 observations in the training sample. For $i \geq 1$, Define

$$X_i^* = \frac{X_i}{S},$$

then the X_i^*'s are from a t distribution with $n_0 - 1$ degrees of freedom. Let $Y_{r,u}^* = \sum_{i=r}^{u} X_i^*$, for $u \geq r \geq 1$, and $Y_{r,u}^* = -\infty$, otherwise. For integers $U \geq R \geq m \geq 2$ and $-\infty < t < \infty$, define

$$N_{R,U}^* = \max_{R \leq k \leq U} \left\{ Y_{k-m+1,k}^* \right\}$$

and

$$\tau_{m,t}^* = \inf \left\{ k \geq m; Y_{k-m+1,k}^* \geq t \right\}.$$

For $0 \leq j \leq m-1$, define $H_{m,t}(j) = 1$. For $M \geq m$, let

$$H_{m,t}(M) = P\left(\tau_{m,t}^* > M\right) = P(N_{m,M}^* < t)$$

and

$$f^*_{m,t}(k) = P\left(\tau^*_{m,t} = k\right).$$

When the values of m and t are clearly understood we will abbreviate $\tau^*_{m,t}$, $H_{m,t}(M)$ and $f^*_{m,t}(k)$ to τ^*, $H(M)$ and $f^*(k)$, respectively.

For $k \geq m$, the random variables

$$\frac{Y^*_{k-m+1,k}}{\sqrt{m}} = \frac{\sum_{i=k-m+1}^{k} X^*_i}{\sqrt{m}} = \frac{\sum_{i=k-m+1}^{k} X_i}{\sqrt{m}S}$$

have a t distribution with $n_0 - 1$ degrees of freedom. The following result holds:

Theorem 2. *For integers $i, m \geq 2$ and $L \geq 1$:*

$$H(M) \geq \frac{H(im)}{\left[1 + \frac{H(Lm-1)-H(Lm)}{H((L+1)m-1)}\right]^{M-im}}, \qquad \text{for } M \geq (i \vee L)m, \tag{2.41}$$

and

$$H(M) \leq H(im)\{1-[H((L+1)m-1)-H((L+1)m)]\}^{M-im}, \qquad \text{for } M \geq (i \vee (L+1))m. \tag{2.42}$$

Proof It follows from the definition of $H(M)$ that

$$\begin{aligned} H(M) &= P(N^*_{m,M} < t) = P\left(\max_{m\le k\le M}\left\{Y^*_{k-m+1,k}\right\} < t\right) \\ &= P\left(\max_{m\le k\le M}\left\{\frac{Y_{k-m+1,k}}{S}\right\} < t\right) = P\left(\max_{m\le k\le M}\left\{\frac{Y_{k-m+1,k}}{\sqrt{m}S}\right\} < \frac{t}{\sqrt{m}}\right) \\ &= P\left(\frac{Y_{1,m}}{\sqrt{m}S} < \frac{t}{\sqrt{m}}, \frac{Y_{2,m+1}}{\sqrt{m}S} < \frac{t}{\sqrt{m}}, ..., \frac{Y_{m+1,M}}{\sqrt{m}S} < \frac{t}{\sqrt{m}}\right), \end{aligned}$$

where for $1 \le j \le M-m+1$, $Y_{j,j+m-1} = \sum_{i=j}^{j+m-1} X_i$ .Therefore,

$$\frac{Y_{1,m}}{\sqrt{m}S}, \frac{Y_{2,m+1}}{\sqrt{m}S}, ..., \frac{Y_{m+1,M}}{\sqrt{m}S}, ...$$

is a sequence of moving sums of t random variables $\{X^*_i = X_i/S; i \ge 1\}$. The inequalities of Theorem 2 follow from Theorem 1.

All the approximations and the inequalities derived in Sections 2.3-2.5 remain valid for the case of σ^2 being unknown by replacing $G(M)$ with $H(M)$.

Remark: For the special case of $i = 2$, $L = 2$ and $M = KM$, $K \ge 4$, we get the following approximation and inequalities for $H(M)$:

$$H(M) \approx H(3m)\left[\frac{H(3m)}{H(2m)}\right]^{K-3} \tag{2.43}$$

$$H(M) \ge \frac{H(2m)}{\left[1+\frac{H(2m-1)-H(2m)}{H(2m-1)H(2m)}\right]^{M-2m}}, \quad \text{for } M \ge 2m. \tag{2.44}$$

For $i = 2$ and $L = 1$, we get

$$H(M) \leq H(2m)\{1 - [H(2m-1) - H(2m)]\}^{M-2m}, \qquad \text{for } M \geq 2m. \tag{2.45}$$

2.7 Numerical Results

In Tables $2.1 - 2.4$, for selected values of parameters M, m and t, approximations and bounds are evaluated for tail probabilities of the statistic defined in Section 2.1, for iid normal random observations with mean 0 and variance 1. These numerical results are obtained from the R algorithm for the multivariate normal distribution in Genz and Bretz (2009). In Table 2.5, for $m = 30$, $M = 750$ and selected values of t, approximations and bounds are presented for scan statistic probabilities, for iid normal random variables with mean 0, unknown variance σ^2. For illustration, the unknown variance σ^2 was estimated from an independent preliminary sample of 50 iid normal random variables with mean 0 and $\sigma^2 = 1$. These numerical results have been evaluated via the R algorithm for the multivariate t distribution in Genz and Bretz (2009).

In Tables $2.1 - 2.4$, $APPRX1$, LB and UB are evaluated via the approximation and bounds for $G(M)$ in (2.15), (2.10) and (2.11), respectively. In Table 2.5, $APPRX1$, LB and UB are evaluated via similar approximations and bounds for $H(M)$. In Tables $2.1 - 2.4$, $APPRX2$ and *Error Bound* are evaluated via Equations (2.16) and (2.17), respectively. In Table 2.5, these quantities are evaluated similarly, for the t distribution

model. In Tables $2.1 - 2.5$,the error bounds have been evaluated only for a restricted range of probabilities, as specified in Section 2.3, based on Haiman (2007), Corollary 2.

In Tables $2.6 - 2.9$, for selected values of parameters, for iid normal observations with mean 0 and variance 1, we present numerical results for the approximation of $E(\tau)$, $SD(\tau)$ (the standard deviation of τ) and their bounds, based on (2.20), (2.27), (2.23), (2.24), (2.29) and (2.31) respectively. In Table 2.10, for selected values of parameters, for iid normal observations with mean 0 and unknown variance, estimated from a preliminary sample of size 50, (for the simulation algorithms without loss of generality we used $\sigma^2 = 1$), we present numerical results for the approximation of $E(\tau)$, $SD(\tau)$ and their bounds.

Based on the numerical results presented in Tables $2.1-2.5$, one can conclude that the approximations and bounds are quite accurate. The approximations for $E(\tau)$ and $SD(\tau)$ in Tables $2.6 - 2.10$ appear to be accurate as well. Their accuracy was confirmed by simulating 10,000 sequences of moving sums of iid normal random variables. Moreover, repeated evaluations of $E(\tau)$ and $SD(\tau)$ via R algorithms in Genz and Bretz (2009) yielded stable results (within 2% of each other).

We have also presented here numerical results for the inequalities of $E(\tau)$ and $SD(\tau)$, derived in Sections 2.4 and 2.5, respectively, but they are unstable for large values of t, when $G(2m)$ (or $H(2m)$) exceeds the value .90. The reason for that is that these approximations include in the denominator the term $G(2m-1)-G(2m)$ (e.g. inequalities for $E(\tau)$ in (2.23) and (2.24)), which is close to 0. Apparently the R algorithm in Genz

and Bretz (2009) is not accurate enough to estimate values of $G(2m-1) - G(2m)$ that are close to 0, thus yielding inaccurate inequalities for $E(\tau)$ and $SD(\tau)$.

2.8 Conclusion

In this chapter approximations and inequalities have been derived for the distribution, expected stopping time and variance of the stopping time associated with moving sums of independent and identically distributed continuous random variables. These approximations and bounds yielded approximations and bounds for scan statistic probabilities that can be employed in detecting a local change in the mean of iid normal observations. Numerical results presented in Section 2.7, have been evaluated via new R algorithms for multivariate normal and t distributions in Genz and Bretz (2009). Based on the numerical results, we can conclude that the approximations and bounds for scan statistic probabilities associated with the distributions of moving sums are accurate and stable, as are the approximations for $E(\tau)$ and $SD(\tau)$. New algorithms are needed for evaluating accurately the inequalities for $E(\tau)$ and $SD(\tau)$, derived in Section 2.4 and 2.5, respectively.

Table 2.1: $P(S_{m,M} \geqslant t)$, M=750, m=30

t	12	14	16	18	19	20	21	22	23	24
LB	.7708	.4901	.2342	.0836	.0334	.0178	.0100	.0037	.0028	.0009
APPRX 1	.7966	.5019	.2358	.0869	.0487	.0263	.0138	.0070	.0033	.0016
APPRX 2	.7852	.4981	.2366	.0867	.0487	.0263	.0138	.0068	.0033	.0014
Error Bound	.	.	.	1.53e-03	4.58e-04	1.26e-04	3.22e-05	7.39e-06	1.52e-06	2.83e-07
UB	.8209	.5130	.2394	.0884	.0656	.0447	.0236	.0078	.0030	.0009

Table 2.2: $P(S_{m,M} \geqslant t)$, M=1500, m=30

t	14	16	18	19	20	21	22	23	24	25
LB	.7473	.4162	.1709	.0691	.0489	.0269	.0138	.0042	.0025	.0028
APPRX 1	.7572	.4213	.1687	.0969	.0530	.0278	.0139	.0066	.0028	.0011
APPRX 2	.7532	.4210	.1702	.0969	.0530	.0278	.0135	.0066	.0028	.0011
Error Bound	.	.	.	9.31e-04	2.59e-04	6.47e-05	1.51e-05	3.12e-06	6.05e-07	1.30e-07
UB	.7705	.4242	.1722	.1223	.0523	.0332	.0164	.0087	.0070	.0028

Table 2.3: $P(S_{m,M} \geqslant t)$, M=1000, m=50

t	14	16	18	20	23	24	25	26	27	28	29
LB	.8361	.6636	.4486	.2530	.0729	.0479	.0388	.0221	.0121	.0074	.0065
APPRX 1	.8732	.6886	.4596	.2601	.0851	.0551	.0350	.0216	.0130	.0077	.0044
APPRX 2	.8535	.6778	.4559	.2587	.0847	.0551	.0349	.0214	.0132	.0077	.0044
Error Bound	.	.	.	.	1.84e-03	7.39e-04	2.84e-04	1.05e-04	3.63e-05	1.21e-05	3.76e-06
UB	.9031	.7148	.4729	.2607	.0886	.0613	.0390	.0252	.0121	.0085	.0065

Table 2.4: $P(S_{m,M} \geqslant t)$, M=2000, m=50

t	17	19	21	23	24	25	26	27	28	29	30
LB	.7952	.5816	.3389	.1699	.1109	.0583	.0414	.0271	.0136	.0056	.0037
APPRX 1	.8249	.5881	.3414	.1658	.1098	.0702	.0436	.0265	.0155	.0089	.0050
APPRX 2	.8188	.5831	.3410	.1655	.1092	.0704	.0436	.0263	.0156	.0091	.0052
Error Bound	.	.	.	.	1.52e-03	5.77e-04	2.14e-04	7.58e-05	2.50e-05	7.68e-06	2.158e-06
UB	.8361	.6009	.3456	.1715	.1115	.0714	.0429	.0271	.0151	.0056	.0037

Table 2.5: $P(S_{m,M} \geqslant t)$, df=49, M=750, m=30

t	12	16	18	19	20	21	22	23	24	25	26
LB	.7973	.2978	.1352	.0766	0.0146	.0056	.0178	.0035	0.0052	0.0047	0.0066
APPRX 1	.8218	.3013	.1348	.0844	.0528	.0321	.0190	.0119	.0062	.0026	.0010
APPRX 2	.8085	.3013	.1341	.0843	.0515	.0321	.0194	.0119	.0057	0026	.0010
Error Bound	.	.	.	1.57e-03	5.72e-04	1.94e-04	6.32e-05	2.27e-05	8.32e-06	4.27e-06	1.01e-06
UB	.8497	.3063	.1369	.1387	.0576	0.1028	.0235	.0726	.0102	.0694	.0444

Table 2.6: Approximations for E(τ) and SD(τ), m=30, M=750

t	12	14	16	18	20	22	24
LB(E(τ))	446.8	1033.5	2676.3	8280.1	19847.0	201527.4	840738.9
E(τ)	479.5	1,066.9	2,724.0	8,003.8	27,181.3	103,398.9	462,279.5
UB(E(τ))	515.7	1103.5	2745.3	8350.8	19895.3	201652.5	840847.8
LB(SD(τ))	367.1	962.7	2610.1	8214.5	19793.1	201435.3	840655.0
SD(τ)	457.3	1,043.5	2,699.6	7,978.5	27,155.3	103,370.9	462,249.0
UB(SD(τ))	560.2	1149.5	2790.1	8398.3	19909.4	201781.5	840952.7

Table 2.7: Approximations for E(τ) and SD(τ), m=30, M=1500

t	14	16	18	19	20	21	22	23	24	25
LB(E(τ))	1027.5	2698.0	7828.9	14352.1	28219.4	45341.5	96796.2	353442.6	390146.8	469960.4
E(τ)	1066.7	2723.5	8010.0	14485.3	27066.7	52747.1	106156.1	222763.9	470894.5	1220988.0
UB(E(τ))	1097.0	2767.5	7895.8	14419.0	28288.7	45398.0	96855.6	353544.8	390195.6	469981.4
LB(SD(τ))	957.0	2631.4	7765.2	14288.7	28155.0	45283.6	96736.9	353362.0	390092.8	469920.3
SD(τ)	1043.4	2699.1	7984.7	14459.6	27040.4	52720.5	106129.0	222736.4	470864.1	1220957.3
UB(SD(τ))	1142.4	2813.2	7937.5	14460.9	28334.2	45424.3	96886.2	353639.6	390210.4	469954.5

Table 2.8: Approximations for E(τ) and SD(τ), m=50, M=1000

t	14	16	18	20	22	24	26	28
LB(E(τ))	453.1	805.8	1539.4	3219.5	6944.6	15090.9	37266.1	128870.5
E(τ)	503.0	863.0	1,601.8	3,228.4	7,080.0	16,934.6	43,943.2	123,016.1
UB(E(τ))	565.5	921.3	1656.0	3338.0	7058.6	15192.5	37360.8	128984.5
LB(SD(τ))	298.9	674.6	1419.8	3105.2	6835.7	14989.7	37169.0	128763.9
SD(τ)	467.1	825.7	1,563.1	3,188.2	7,038.6	16,892.2	43,899.7	122,970.7
UB(SD(τ))	636.7	995.3	1731.7	3416.7	7130.9	15246.3	37404.3	129056.9

Table 2.9: Approximations for E(τ) and SD(τ), m=50, M=2000

t	17	19	21	23	24	25	26	27	28	29	30
LB(E(τ))	1112.6	2174.3	4660.9	10444.9	16590.1	26446.2	46283.3	71075.4	143402.0	119649.1	192425.2
E(τ)	1164.2	2250.5	4733.4	10844.8	16876.2	26871.8	43960.2	73681.8	125503.3	218766.7	389484.8
UB(E(τ))	1229.9	2290.3	4776.6	10555.9	16701.9	26556.6	46401.1	71183.0	143528.8	119708.7	192477.3
LB(SD(τ))	987.5	2058.8	4549.8	10338.3	16483.7	26340.9	46174.6	70971.8	143289.0	119569.8	192349.6
SD(τ)	1126.3	2211.1	4692.7	10803.0	16833.5	26828.3	43916.5	73638.1	125459.1	218722.0	389439.1
UB(SD(τ))	1306.3	2365.1	4851.4	10623.8	16771.1	26623.6	46479.1	71246.0	143620.4	119699.5	192457.1

Table 2.10: Approximations for E(τ) and SD(τ), m=30, M=750, df=49

t	12	16	18	20	22	24
LB(E(τ))	407.6	2010.3	4954.9	12203.5	30258.7	69679.0
E(τ)	444.0	2,051.5	4,955.0	13,412.0	37,887.9	116,889.6
UB(E(τ))	477.1	2081.2	5024.3	12266.5	30312.6	69719.7
LB(SD(τ))	325.6	1942.2	4889.5	12142.0	30202.2	69629.2
SD(τ)	422.8	2,028.2	5,014.5	13,387.2	37,862.4	116,864.7
UB(SD(τ))	522.5	2128.7	5069.7	12302.6	30334.8	69722.1

Chapter 3

Two Dimensional Scan Statistics

3.1 Introduction

For $1 \leq i \leq M_1$ and $1 \leq j \leq M_2$, let X_{ij} be iid observations from a continuous distribution with mean μ and variance σ^2. For $2 \leq m_k \leq M_k - 1, 1 \leq i_k \leq M_k - m_k + 1$ and $k = 1, 2$, define

$$Y_{i_1,i_2} = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} X_{ij}, \tag{3.1}$$

as a moving sum in a $m_1 \times m_2$ rectangular grid. These moving sums follow a multivariate distribution with a $(M_1 - m_1 + 1) \times (M_2 - m_2 + 1)$ dimensional mean vector $(m_1 m_2 \mu, ..., m_1 m_2 \mu)'$ and covariance matrix Σ whose elements equal to σ^2 times the number of observations that belong to both of the corresponding moving sums. A two dimensional scan statistic, $S_{m_1,m_2}(M_1, M_2)$, is defined as the maximum of moving sums over all $m_1 \times m_2$ rectangular grids within the range to be monitored:

$$S_{m_1,m_2}(M_1, M_2) = \max \left\{ Y_{i_1,i_2}; 1 \leq i_k \leq M_k - m_k + 1, k = 1, 2 \right\}. \tag{3.2}$$

To simplify the presentation of results, we will assume that $M_1 = M_2 = M$, $m_1 = m_2 = m$ and $M = Lm$, where m, $L \geq 3$. For $2 \leq m \leq M$ and $-\infty < t < \infty$, let

$$G_{m,t}(M) = P(\max\{Y_{i_1,i_2}; 1 \leq i_k \leq M_k - m_k + 1, k = 1, 2\} < t). \tag{3.3}$$

The distribution of the scan statistic $S_{m,m}(M, M)$ is given by

$$P(S_{m,m}(M, M) < t) = G_{m,t}(M).$$

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.

The chapter is organized as follows. Approximations and inequalities for the distribution of two dimensional scan statistics, based on iid observations from a continuous distribution, are given in Section 3.2 and 3.3 respectively. In Section 3.4, we present one and two dimensional multiple window scan statistics for normal data. Models with known and unknown mean and variance are discussed. These scan statistics are designed for detecting a local change in the mean of a normal model in a given region. Based on Genz and Bretz (2009) algorithms for the multivariate normal and t distributions, we present algorithms for implementing these multiple window scan statistics and evaluating their power for a specified alternative. In Section 3.5, for selected values of the parameters, numerical results are presented to evaluate the performance of proposed

approximations, inequalities and multiple window scan statistics. Concluding remarks are presented in Section 3.6.

3.2 Approximations for $G(M)$

For $1 \leq i_1, i_2 \leq M - m + 1$, define the event

$$A_{i_1,i_2} = (Y_{i_1,i_2} \geq t).$$

Then,

$$P(S_{m,m} \geq t) = P(\bigcup_{i_1=1}^{M-m+1} \bigcup_{i_2=1}^{M-m+1} A_{i_1,i_2}).$$

To derive a Markov-like product-type approximation for $G(M)$, let

$$B_{i_1} = (\bigcap_{i_2=1}^{M-m+1} A^c_{i_1,i_2}),$$

where $1 \leq i_1 \leq M - m + 1$. Then

$$G(M) = P(B_1) \prod_{i_1=2}^{M-m+1} P(B_{i_1}|B_{i_1-1} \bigcap ... \bigcap B_1).$$

To approximate $G(M)$, we employ the following approximation suggested in Naus and Sheng (1996):

$$P(B_{i_1}|B_{i_1-1} \bigcap ... \bigcap B_1) \approx P(B_{i_1}|B_{i_1-1}).$$

Therefore,

$$G(M) \approx P(B_1) \prod_{i_1=2}^{M-m+1} P(B_{i_1}|B_{i_1-1}) = \frac{[P(B_1 \bigcap B_2)]^{M-m}}{[P(B_1)]^{M-m-1}}. \tag{3.4}$$

To evaluate $P(B_1)$ and $P(B_1 \bigcap B_2)$ via Genz and Bretz (2009) R algorithm, we need to evaluate a $2(M-m+1)$ dimensional multivariate probability. Since the Genz and Bretz (2009) R algorithm is only valid for dimensions lower than 1000, in some applications the above approximation can not be used. We propose to employ the following approximation (Boutsikas and Koutras, 2000), when the dimension exceeds 1000:

$$G(M) \approx \frac{[P(S_{m,m}(m+1,m+1) < t)]^{(M-m)^2}}{[P(S_{m,m}(m,m+1) < t)]^{(M-m-1)(M-m)}} \times \frac{q_{m,2m-1}^{(M-2m)(M-m-1)}}{q_{m,2m}^{(M-2m+1)(M-m-1)}}, \tag{3.5}$$

where

$$P(S_{m,m}(m+1,m+1) < t) = P(A_{1,1}^c \bigcap A_{1,2}^c \bigcap A_{2,1}^c \bigcap A_{2,2}^c),$$

$$P(S_{m,m}(m,m+1) < t) = q_{m,m+1},$$

and for $m \leq l \leq 2m$

$$q_{m,l} = P(\bigcap_{i=1}^{l-m+1} A_{1,i}^c).$$

Haiman and Preda (2006) also derived accurate approximations for $G(M)$ for the case of two dimensional iid discrete random variables. These approximations are valid as well for iid continuous random variables. A nice feature of these approximations is that a sharp error bound can be easily evaluated. For the problem at hand, for any t and $M \geq 3m$, such that $1 - P(S_{m,m}(2m, M) < t) \leq 0.025$, the following approximation for $G(M)$ is obtained:

$$G(M) \approx 2(Q_2 - Q_3)[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{-L+1}, \tag{3.6}$$

where

$$Q_2 \approx 2(q^*_{2m,2m} - q^*_{2m,3m})[1 + q^*_{2m,2m} - q^*_{2m,3m} + 2(q^*_{2m,2m} - q^*_{2m,3m})^2]^{-L+1},$$

$$Q_3 \approx 2(q^*_{3m,2m} - q^*_{3m,3m})[1 + q^*_{3m,2m} - q^*_{3m,3m} + 2(q^*_{3m,2m} - q^*_{3m,3m})^2]^{-L+1},$$

and $P(S_{m,m}(am, bm) < t) = q^*_{am,bm}$, which stands for the distribution of scan statistics defined in a am by bm rectangular region. An error bound for $G(M)$ is given by:

$$E = E_{app} + E_{sim} \tag{3.7}$$

where E_{app} arises from the approximation process in (3.6), and E_{sim} arises from the simulation process of $q^*_{2m,2m}$, $q^*_{2m,3m}$, $q^*_{3m,2m}$ and $q^*_{3m,3m}$. These two error terms can be

found in Haiman and Preda (2006). In Section 3.4, based on approximations (3.4), (3.5) and (3.6), we will evaluate the performance of these approximations for the distribution of $S_{m,m}$.

3.3 Inequalities for $G(M)$

We now proceed to derive second order Bonferroni-type inequalities for $G(M) = 1 - P(S_{m,m} \geq t)$. Since,

$$\begin{aligned} P(S_{m,m} \geq t) &= P(\bigcup_{i_1=1}^{M-m+1} \bigcup_{i_2=1}^{M-m+1} A_{i_1,i_2}) \\ &= P(\bigcup_{i_1=1}^{M-m+1} B_{i_1}^c), \end{aligned}$$

we will derive Bonferroni-type inequalities in terms of the events B_{i_1}, where $1 \leq i_1 \leq M - m + 1$. It follows from Hunter (1976) that

$$\begin{aligned} P(\bigcup_{i_1=1}^{M-m+1} B_{i_1}^c) &\leq \sum_{i_1=1}^{M-m+1} P(B_{i_1}^c) - \sum_{i_1=1}^{M-m} P(B_{i_1}^c \bigcap B_{i_1+1}^c) \\ &= (M-m+1)P(B_1^c) - (M-m)P(B_1^c \bigcap B_2^c). \end{aligned}$$

Substitute

$$P(B_1^c) = 1 - P(B_1)$$

and

$$P(B_1^c \bigcap B_2^c) = 1 - P(B_1 \bigcup B_2) = 1 - 2P(B_1) + P(B_1 \bigcap B_2)$$

in the above inequality to get

$$P(\bigcup_{i_1=1}^{M-m+1} B_{i_1}^c) \leq 1 - (M-m)P(B_1 \bigcap B_2) + (M-m-1)P(B_1).$$

We get the following lower bound:

$$G(M) \geq (M-m)P(B_1 \bigcap B_2) - (M-m-1)P(B_1). \tag{3.8}$$

To derive an upper bound for $G(M)$, we employ the inequality from Kwerel (1975) to get

$$G(M) \leq 1 - \frac{2s_1}{b} + \frac{2s_2}{b(b-1)}, \tag{3.9}$$

where b is integer part of $2 + 2s_2/s_1$,

$$s_1 = (M-m+1)(1-P(B_1))$$

and

$$s_2 = \sum_{j=2}^{M-m+1}\sum_{i=1}^{j-1} P(B_i^c \bigcap B_j^c) = \sum_{j=2}^{M-m+1}\sum_{i=1}^{j-1}[1 - 2P(B_1) + P(B_i \bigcap B_j)]$$

$$= 0.5(M-m+1)(M-m)[1-2P(B_1)] + \sum_{j=2}^{M-m+1}\sum_{i=1}^{j-1} P(B_i \bigcap B_j)$$

$$= 0.5(M-m+1)(M-m)[1-2P(B_1)] + \sum_{j=2}^{m}(M-m+2-j)P(B_1 \bigcap B_j)$$

$$+ 0.5(M-2m+1)(M-2m+2)[P(B_1)]^2.$$

The last equality follows from the fact that for $j - i \geq m$, the events B_i and B_j are independent and therefore $P(B_i \bigcap B_j) = [P(B_1)]^2$.

In Section 3.4, one and two dimensional multiple window scan statistics will be developed and evaluated for normal data, based on approximation (2.15) and (3.4), respectively.

3.4 Multiple Window Scan Statistics

We now introduce a multiple window scan statistic for one dimensional normal data. This scan statistic can be used in detecting a change of size m in the mean within a sequence of M observations via testing the null hypothesis, H_0, that assumes $X_i, 1 \leq i \leq M$, are iid normal random variables with mean $\mu = \mu_0$ and variance σ^2. For the alternative hypothesis, H_1, one assumes a local change in mean within a segment of m

consecutive observations

$$S(i_0, m) = \{i_0, i_0 + 1, ..., i_0 + m - 1\},$$

where both the window size m, $2 \leq m \leq M/4$ and the location $i_0, 1 \leq i_0 \leq M-m+1$, are unknown. Under H_1, for any $i_0 \leq i \leq i_0+m-1$, X_i has a normal distribution with mean $\mu = \mu_1$ and variance σ^2, where $\mu_1 > \mu_0$. For $i \notin S(i_0, m)$, X_i is distributed according to the distribution specified by the null hypothesis. When the length of the sliding window m is known, the generalized likelihood ratio test rejects the null hypothesis in favor of the local change in alternative hypothesis when S_m exceeds the a threshold t, where t is determined from $P(S_m \geq t|H_0) = \alpha$, and α is a pre-specified significance level of the testing procedure (Glaz et al., 2001, chap. 13). Here we assume that the window size m is unknown. Both cases of variance σ^2 known and unknown will be considered. Since the size of the sliding window m is unknown, we propose to investigate the performance of a multiple window scan statistic based on a sequence of n fixed window scan statistics: $S_{m_1}, ..., S_{m_n}$, where $2 \leq m_1 < m_2 < .... < m_n \leq M/4$, where the lengths of the n sliding windows, are chosen by the experimenter. For $1 \leq j \leq n$, let t_j be the observed value of S_{m_j} and $p_j = P(S_{m_j} \geq t_j \mid H_0)$ the associated p-value. To test H_0 vs H_1 we propose the following test statistic:

$$P^{(1)}_{\min} = \min\{p_j; 1 \leq j \leq n\}, \tag{3.10}$$

the *minimum P-value statistic*. In the context of multiple window scan statistics, $P^{(1)}_{\min}$ has been introduced in Hoh and Ott (2000) for a one dimensional $0-1$ iid Bernoulli model. It has been extended to two-dimensional multiple window scan statistics for binomial and Poisson data in Zhang and Glaz (2008) and Chen and Glaz (2009).

Since the exact distribution for the $P^{(1)}_{min}$ statistic is unknown, for a given significant level α, the critical value p_α,

$$P_{H_0}\left(P^{(1)}_{min} \leq p_\alpha\right) = \alpha,$$

has to be evaluated via simulation. In each run of the simulation, we generate M observations under the null hypothesis. Then we scan the whole region with multiple moving windows of sizes m_1, m_2, ... and m_n, and record the observed values of the fixed window scan statistics, $S_{m_1}, ..., S_{m_n}$, denoted by $t_1, t_2, ..., t_n$, respectively. At the next stage, a randomized quasi Monte-Carlo R algorithm is employed to evaluate the observed p values $p_j = P(S_{m_j} \geq t_j \mid H_0)$, $1 \leq j \leq n$. The minimum of these p values is recorded and then the whole process is repeated N times. p_α will be calculated as the $\alpha * 100$ percentile of the simulated distribution of $P^{(1)}_{\min}$ statistic.

We now introduce a multiple window scan statistic for two dimensional normal data. This statistic can be used to detect a local change of mean within a m_1 by m_2 grid by testing the null hypothesis, H_0, that assumes $\{X_{ij}\}$ are iid observations from a normal distribution with mean $\mu = \mu_0$ and variance σ^2. For the alternative hypothesis, H_1, of

a local change in the mean, one often specifies a rectangular subregion

$$S(i_1, i_2) = [i_1, i_1 + m_1 - 1] \times [i_2, i_2 + m_2 - 1],$$

where X_{ij} has a normal distribution with mean $\mu = \mu_1$, where $\mu_1 > \mu_0$. For $i, j \notin S(i_1, i_2)$, X_{ij} is distributed according to the distribution specified by the null hypothesis. When the size of the scanning $m_1 \times m_2$ window is known, the generalized likelihood ratio test rejects the null hypothesis in favor of the local increase in mean from alternative hypothesis H_1, when S_{m_1,m_2} exceeds a threshold t, where t is determined from $P(S_{m_1,m_2} \geq t|H_0) = \alpha$, and α is a specified significance level of the testing procedure (Glaz et al., 2001, chap. 16.1). The use of S_{m_1,m_2} for testing the null hypothesis of randomness for binomial and Poisson data has been of interest in many areas of science and technology (Glaz et al., 2001, chap. 16.1).

In what follows, to simplify the presentation of the results, we will assume that $M_1 = M_2 = M$ and $m_1 = m_2 = m$. Let $1 \leq i_0, j_0 \leq M - m + 1$ and $2 \leq m \leq M/4$ be unknown parameters. Since the size of the window where a change happens is unknown, we propose to investigate the performance of a multiple window scan statistic based on a sequence of n fixed window scan statistics: $S_{m_1,m_1}, ..., S_{m_n,m_n}$, where $2 \leq m_j < m_{j+1} \leq M/4$, $1 \leq j \leq n-1$. For $1 \leq j \leq n$, let t_j be the observed value of S_{m_j,m_j} and $p_j = P(S_{m_j,m_j} \geq t_j \mid H_0)$ be the associated p-value. To test H_0 vs. H_1 we propose

to investigate the performance of the test statistic:

$$P_{\min}^{(2)} = \min\{p_j; 1 \leq j \leq n\}. \tag{3.11}$$

This two-dimensional multiple window scan statistic is of interest in the area of environmental sciences (Patil et al., 2009) and signal detection in a sensor network (Guerriero et al., 2009).

To implement the $P_{\min}^{(2)}$ statistic, approximation (3.4) derived in Section 3.2 is employed. Similar to the one dimensional case, for a given significance level α, the critical value p_α,

$$P_{H_0}\left(P_{min}^{(2)} \leq p_\alpha\right) = \alpha,$$

has to be evaluated via simulation. In each run of the simulation, we generate M by M observations under the null hypothesis. Then, we scan the whole region with multiple two dimensional moving windows of sizes $m_1 \times m_1, m_2 \times m_2$, ..., and $m_n \times m_n$, and record the observed values of the fixed window scan statistics $S_{m_1,m_1}, ..., S_{m_n,m_n}$, denoted by $t_1, t_2, ..., t_n$, respectively. At the next stage, a randomized quasi Monte-Carlo R algorithm is employed to evaluate the observed p values: $p_j = P(S_{m_j,m_j} \geq t_j \mid H_0)$, $1 \leq j \leq n$, via approximation (3.4). The minimum of these p values is recorded and then the whole process is repeated N times. p_α will be calculated as the $\alpha * 100$ percentile of the simulated distribution of $P_{\min}^{(2)}$ statistic.

In this section we are interested in investigating the performance of multiple window

scan statistics for one and two dimensional normal data. Implementation of these scan statistics are based on approximations for the fixed window scan statistics. For one dimensional data, section 2.6 discussed the cases when the mean μ and variance σ^2 are unknown. It was shown there that without loss of generality one can always assume that $\mu = 0$. When σ^2 is unknown, it can be estimated by the sample variance S^2 from a training sample of size n_0, independent of the observed data. A larger value of n_0 will yield a more accurate estimator for σ^2. In this case the following sequence is considered:

$$X_i^* = \frac{X_i}{S}, i = 1, ..., M, \tag{3.12}$$

where $\{X_i; i = 1, ..., M\}$, are iid normal with $\mu = 0$ and unknown variance σ^2. $Y^*_{j-m+1,j}$ is defined accordingly as $\frac{Y_{j-m+1,j}}{S}$. The sequence of random variables $\left\{\frac{X_i}{\sigma}\right\}$ follow a multivariate normal distribution with a zero mean vector and identity covariance matrix, and since $\frac{(n_0-1)S^2}{\sigma^2}$ follows a chi-square distribution with degrees of freedom $n_0 - 1$, independent of the data, thus we have $\{X_i^*\}$ follow a multivariate t distribution with degrees of freedom $n_0 - 1$ and identity covariance matrix. We now proceed to derive the distribution of $\left\{Y^*_{j-m+1,j}\right\}$. It can be shown that $\left\{\frac{Y_{j-m+1,j}}{\sigma}\right\}$ follow a multivariate normal distribution with a zero mean vector and a covariance matrix whose elements equal to the number of observations that belong to both of the corresponding moving sums. Since $\frac{(n_0-1)S^2}{\sigma^2}$ follows a chi-square distribution with degrees of freedom $n_0 - 1$, independent of the data, we have that $\left\{Y^*_{j-m+1,j}\right\}$ follow a multivariate t distribution with degrees

of freedom $n_0 - 1$ and the same covariance matrix. We can employ the randomized quasi Monte-Carlo R algorithm established by Genz and Bretz (2009) to evaluate the multivariate t probabilities for the distribution of fixed window scan statistic. See section 2.6 for more details.

For normal observations in a two dimensional grid, without loss of generality one can always assume that $\mu = 0$. Otherwise, one can extend the approach in Bauer and Hackl (1978) and consider the following sequence of 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 iid normal random variables with mean 0 and variance σ^2. If σ^2 is known, without loss of generality we can assume $\sigma^2 = 1$. We employ a randomized quasi Monte-Carlo R algorithm for evaluating multivariate normal probabilities (Genz and Bretz, 2009) to evaluate the approximations and inequalities for the distribution of fixed window scan statistic derived in the previous section.

If the variance σ^2 is unknown, assume that a training sample of n_0 iid normal random variables, independent of the $\{X_{ij}\}$, and with the same distribution has been observed. Let S^2 be the sample variance based on n_0 observations in the training sample, independent of the observed data. Consider the following sequence:

$$X^*_{ij} = \frac{X_{ij}}{S}, i, j = 1, ..., M, \tag{3.13}$$

and $Y^*_{i_1,i_2}$, $S^*_{m,m}$ and $G^*(M)$ will be defined accordingly, with X_{ij} replaced by X^*_{ij}. Similar to one dimensional case discussed above, the sequence of random variables $\left\{X^*_{ij}\right\}$ follow a multivariate t distribution with degrees of freedom $n_0 - 1$ and identity covariance matrix, and $\left\{Y^*_{i_1,i_2}\right\}$ follow a multivariate t distribution with $n_0 - 1$ degrees of freedom and a covariance matrix whose elements equal to the number of observations that belong to both of the corresponding moving sums. It follows that the approximations and inequalities derived in Sections 3.2 and 3.3 hold for $G^*(M)$. Therefore, we can employ a randomized quasi Monte-Carlo R algorithm for evaluating multivariate t probabilities (Genz and Bretz, 2009) to evaluate approximations and inequalities for the distribution of a fixed window two dimensional scan statistic, when variance σ^2 is unknown.

To evaluate the performance of multiple window scan statistics, $P^{(1)}_{min}$ and $P^{(2)}_{min}$, and compare it with that of fixed window scan statistics, in detecting a change of the mean within a window of unknown size, we designed the following algorithms to evaluate the power of these scan statistics. In these algorithms, we have a specified number and length of the moving windows. For the alternative hypothesis we pre-selected a locally increased mean $\mu_1 > 0$. For a specified significance level α, we have evaluated the power under each alternative hypothesis. The following algorithms are used to evaluate the power. First, we present the steps for the one dimensional case, with σ^2 being both known and unknown.

1. If σ^2 is known, generate M observations from a normal distribution with $\mu_0 = 0$ and $\sigma^2 = 1$, and in a specified window of length m replace them with observations

with mean $\mu_1 > 0$.

If σ^2 is unknown, by using a learning sample of size n_0, generate M observations from a multivariate central-t distribution (3.12) with $n_0 - 1$ degrees of freedom and an identity covariance matrix, and in a specified window of length m replace them with observations with mean $\mu_1 > 0$.

2. Scan the whole region with selected window sizes, and let $t_1, ..., t_n$ be the observed values of the fixed window scan statistics $S_{m_1}, ..., S_{m_n}$, respectively.

3. For a fixed window of length m_j and a specified significance level α, evaluate $p_j = 1 - P(S_{m_j} < t_j)$, and reject H_0 if $p_j < \alpha$.

4. For the multiple window scan statistic, $P_{min}^{(1)} = min\{p_j; 1 \leq j \leq n\}$, reject H_0 if $P_{min}^{(1)} < p_\alpha$.

5. Repeat steps 1-4 N times and count how many times out of N trials, we have rejected H_0 with both the fixed and multiple window scan statistics.

For the two dimensional case a similar algorithm is employed.

1. If σ^2 is known, we generate M by M observations from a normal distribution with $\mu_0 = 0$ and $\sigma^2 = 1$, and in a specified $m \times m$ rectangular window replace them with observations with mean $\mu_1 > 0$.

   If σ^2 is unknown, by using a learning sample of size n_0, generate M by M observations from a multivariate central-t distribution (3.13) with $n_0 - 1$ degrees of

freedom and an identity covariance matrix, and in a specified $m \times m$ rectangular window replace them with observations with $\mu_1 > 0$.

2. Scan the whole region with selected window sizes, and let $t_1, ..., t_n$ be observed fixed window scan statistics, respectively.

3. For a fixed window of length m_j and significance level α, evaluate $p_j = 1 - P(S_{m_j,m_j} < t_j)$. Reject the null hypothesis if $p_j < \alpha$.

4. For the multiple window scan statistics, $P^{(2)}_{min} = min\{p_j; 1 \leq j \leq n\}$, we reject the null hypothesis if $P^{(2)}_{min} < p_\alpha$.

5. Repeat steps 1-4 N times and count how many times out of N we reject the null hypothesis for both variable and fixed window scan statistics.

3.5 Numerical Results

In this section numerical results are presented to evaluate, for selected values of the parameters, the accuracy of approximations and inequalities for the distribution of two dimensional scan statistics, and to compare the power of multiple window scan statistics with fixed window scan statistics. In Table 3.1, for $M = 250$, $m = 10$ and selected threshold values t, approximations and inequalities are evaluated for tail probabilities $1 - G(M)$, defined in Sections 3.2 and 3.3, for iid normal random observations with mean 0 and variance 1. These numerical results are obtained from the R algorithm

for the multivariate normal distribution in Genz and Bretz (2009). In Tables 3.2 – 3.3, for $M = 250$, $m = 10$ and selected values of threshold t, approximations and inequalities are presented for tail probabilities for iid normal random variables with mean 0, unknown variance σ^2. For illustration, the unknown variance σ^2 was estimated from an independent preliminary sample of 16 and 101 iid normal random variables. These numerical results are evaluated via the R algorithm for the multivariate t distribution in Genz and Bretz (2009).

In Table 3.1, LB, $Appx1$, $Appx2$, $Appx3$, $Error$, $Error\ App$, $Error\ Sim$ and UB are evaluated via the approximations and inequalities for $1-G(M)$ in (3.9), (3.4), (3.5), (3.6), (3.7) and (3.8) respectively. In Haiman and Preda (2006) approach, we used 10^7 simulation runs to estimate $q^*_{2m,2m}$, $q^*_{2m,3m}$, $q^*_{3m,2m}$ and $q^*_{3m,3m}$. The error bounds have been evaluated only for a restricted range of probabilities, as specified in Section 3.2, based on Haiman and Preda (2006). In Tables 3.2 – 3.3, these quantities are evaluated similarly for the t distribution via the same formula but with respect to X^*_{ij} instead of X_{ij}. Haiman and Preda (2006) approximation is not valid for the multivariate t distribution used here, as the sequence $\{X^*_{ij}\}$ is not independent. Conditional on $\frac{(n_0-1)S^2}{\sigma^2}$, one can evaluate all the components that are needed to evaluate the approximation in Haiman and Preda (2006).

In Tables 3.4 – 3.9, numerical results are presented to evaluate the accuracy of achieving a specified probability of Type I error for multiple window scan statistics, and to compare their power with fixed window scan statistics. For a given significance level

$\alpha = 0.05$, the critical values p_α are evaluated via simulations with $10,000$ trials using the algorithms in Section 3.4. For selected values of parameters, we used $1,000$ replications to simulate the power of $P_{min}^{(1)}$ and $P_{min}^{(2)}$ statistics and compare it with the power of fixed window scan statistics.

In tables $3.4 - 3.6$, numerical results are presented for the one dimensional case. In Table 3.4, based on a simulation with $1,000$ trials, we evaluate the power of $P_{min}^{(1)}$ for $M = 250$ iid normal observation with mean μ and variance $\sigma^2 = 1$, where the mean parameter under the null hypothesis $\mu = 0$, and the mean parameter $\mu = \mu_1 > 0$ within a consecutive sequence of m observations, under the alternative hypothesis. The power of $P_{min}^{(1)}$ is compared to that of the fixed window scan statistics S_{m_j} where the length of the scanning window is m_j. In Tables $3.5 - 3.6$, under the same setup, based on a simulation with $1,000$ trials, we evaluate the power of $P_{min}^{(1)}$, for iid normal random variables with unknown variance σ^2, with mean $\mu = 0$ under the null hypothesis and $\mu = \mu_1 > 0$ within a consecutive sequence of m observations, under the alternative hypothesis. For illustration, the unknown variance σ^2 was estimated from an independent preliminary sample of 16 and 101 iid normal observations.

In tables $3.7 - 3.9$, numerical results are presented for two dimensional case. In table 3.7, for $M = 250$, with mean parameter under null hypothesis $\mu_0 = 0$, selected $\mu = \mu_1 > 0$ under alternative hypothesis, and a local change size m under alternative hypothesis, the power of $P_{min}^{(2)}$ is evaluated via 1000 simulations and compared with fixed window scan statistics S_{m_j,m_j} where scanning window is fixed at m_j by m_j, for

iid normal random observations with mean 0 and variance 1. In tables $3.8 - 3.9$, for $M = 50$, under the same setup, the power of $P_{min}^{(2)}$ is evaluated and compared with fixed window scan statistics for iid normal random variables with mean 0, unknown variance σ^2. The unknown variance σ^2 was estimated from an independent preliminary sample of 16 and 101 iid normal random variables, respectively.

Based on the numerical results presented in Tables $3.1 - 3.3$, one can conclude that the approximations and inequalities are quite accurate. By comparing the numerical results for power calculations in Tables $3.4 - 3.9$, fixed window scan statistics with correctly specified window size turned out to be most powerful. The multiple window scan statistics were slightly less powerful in that case, but they outperformed the fixed window scan statistics with an incorrectly specified window size where a change in mean has occurred.

3.6 Conclusion

In this chapter, we investigated the performance of multiple window scan statistics for iid normal data in one and two dimensional regions. Based on the numerical results it is evident that, when the size of the region where a change in the mean has occurred is unknown, the multiple window scan statistics outperform the fixed window scan statistics. Numerical results presented in Section 3.5, have been evaluated via new R algorithms for multivariate normal and t distributions in Genz and Bretz (2009). To implement

these multiple window scan statistics for other models of iid continuous observations, new effective algorithms for general multivariate distributions need to be developed.

Table 3.1: two dimensional Normal variables, M=250, m=10

t	LB	Appx1	Appx2	Appx3	Error Bound	Error App	Error Sim	UB
45	0.0603	0.0982	0.0829	0.0829	0.0061	0.0018	0.0043	0.1033
46	0.0456	0.0617	0.0539	0.0575	0.0044	0.0008	0.0036	0.0637
47	0.0080	0.0358	0.0487	0.0361	0.0031	0.0003	0.0028	0.0364
48	0.0060	0.0167	0.0287	0.0242	0.0025	0.0001	0.0023	0.0169
49	0.0020	0.0131	0.0221	0.0142	0.0018	0.0000	0.0018	0.0131
50	0.0011	0.0068	0.0082	0.0084	0.0014	0.0000	0.0014	0.0068

Table 3.2: two dimensional t variables, df=15, M=250, m=10

t	Lower Bound	Appx1	Appx2	Upper Bound
63	0.0441	0.0811	0.1447	0.0846
64	0.0575	0.0804	0.1260	0.0838
65	0.0442	0.0856	0.0989	0.0894
66	0.0331	0.0763	0.0775	0.0793
67	0.0165	0.0435	0.0752	0.0445
68	0.0125	0.0355	0.0714	0.0362

Table 3.3: two dimensional t variables, df=100, M=250, m=10

t	Lower Bound	Appx1	Appx2	Upper Bound
47	0.0388	0.0938	0.1046	0.0985
48	0.0564	0.0800	0.0586	0.0833
49	0.0143	0.0352	0.0607	0.0359
50	0.0186	0.0288	0.0334	0.0292
51	0.0109	0.0246	0.0162	0.0250

Table 3.4: one dimensional Normal variables, $\mu_0 = 0$, M=250

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.5	0.091	0.074	0.100	0.076	0.061	0.067
	1	0.424	0.339	0.468	0.317	0.228	0.189
	1.5	0.909	0.828	0.926	0.770	0.589	0.465
15	0.5	0.150	0.116	0.149	0.155	0.119	0.099
	1	0.742	0.506	0.695	0.773	0.618	0.505
	1.5	0.993	0.932	0.987	0.994	0.971	0.921
20	0.5	0.247	0.153	0.209	0.263	0.264	0.189
	1	0.895	0.602	0.819	0.878	0.913	0.834
	1.5	1.000	0.981	0.997	1.000	1.000	1.000
Type I Error		0.050	0.045	0.048	0.047	0.046	0.063

Table 3.5: one dimensional t variables, df=15, $\mu_0 = 0$, M=250

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.5	0.064	0.042	0.056	0.053	0.066	0.066
	1	0.148	0.105	0.186	0.107	0.093	0.083
	1.5	0.559	0.346	0.657	0.390	0.277	0.217
15	0.5	0.067	0.053	0.067	0.060	0.060	0.059
	1	0.373	0.148	0.322	0.451	0.320	0.253
	1.5	0.898	0.510	0.844	0.930	0.821	0.710
20	0.5	0.111	0.062	0.085	0.106	0.129	0.106
	1	0.630	0.183	0.456	0.618	0.700	0.588
	1.5	0.983	0.633	0.941	0.978	0.988	0.967
Type I Error		0.052	0.024	0.033	0.027	0.028	0.027

Table 3.6: one dimensional t variables, df=100, $\mu_0 = 0$, M=250

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.5	0.089	0.090	0.106	0.074	0.049	0.052
	1	0.338	0.288	0.427	0.264	0.188	0.152
	1.5	0.866	0.770	0.911	0.735	0.555	0.442
15	0.5	0.126	0.101	0.138	0.138	0.109	0.103
	1	0.666	0.426	0.646	0.731	0.567	0.449
	1.5	0.990	0.913	0.986	0.996	0.954	0.902
20	0.5	0.211	0.115	0.173	0.228	0.248	0.201
	1	0.858	0.507	0.786	0.863	0.898	0.814
	1.5	1.000	0.964	0.999	0.999	1.000	1.000
Type I Error		0.050	0.045	0.053	0.038	0.044	0.043

Table 3.7: two dimensional Normal variables, $\mu_0 = 0$, M=250

m	μ_1	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{15,15}$	$S_{20,20}$	$S_{25,25}$
10	0.2	0.056	0.044	0.061	0.048	0.043	0.039
	0.4	0.269	0.127	0.354	0.087	0.046	0.037
	0.5	0.606	0.219	0.716	0.159	0.100	0.059
15	0.2	0.714	0.172	0.605	0.741	0.441	0.222
	0.4	0.824	0.171	0.675	0.860	0.516	0.254
	0.5	0.989	0.465	0.964	0.990	0.890	0.584
20	0.1	0.073	0.056	0.068	0.056	0.062	0.048
	0.2	0.209	0.036	0.095	0.207	0.290	0.126
	0.4	1.000	0.326	0.945	0.998	1.000	0.985
Type I Error		0.050	0.052	0.039	0.041	0.038	0.036

Table 3.8: two dimensional t variables, df=15, $\mu_0 = 0$, M=50

m	μ_1	P_{min}	$S_{5,5}$	$S_{7,7}$	$S_{10,10}$	$S_{15,15}$
5	0.5	0.049	0.041	0.030	0.040	0.020
	0.8	0.099	0.082	0.069	0.080	0.055
	1	0.263	0.168	0.196	0.148	0.154
7	0.5	0.318	0.184	0.178	0.210	0.189
	0.8	0.502	0.369	0.334	0.383	0.360
	1	0.718	0.589	0.588	0.553	0.558
10	0.5	0.472	0.317	0.336	0.313	0.323
	0.8	0.967	0.818	0.821	0.847	0.839
	1	1.000	0.994	0.984	0.978	1.000
Type I Error		0.050	0.034	0.034	0.034	0.035

Table 3.9: two dimensional t variables, df=100, $\mu_0 = 0$, M=50

m	μ_1	P_{min}	$S_{5,5}$	$S_{7,7}$	$S_{10,10}$	$S_{15,15}$
5	0.5	0.077	0.057	0.064	0.056	0.066
	0.8	0.298	0.196	0.204	0.159	0.165
	1	0.425	0.324	0.319	0.253	0.286
7	0.5	0.309	0.179	0.173	0.204	0.184
	0.8	0.502	0.374	0.375	0.366	0.347
	1	0.830	0.691	0.691	0.623	0.617
10	0.3	0.260	0.164	0.157	0.133	0.169
	0.5	0.835	0.561	0.615	0.536	0.604
	0.8	1.000	0.977	0.989	0.992	0.966
Type I Error		0.048	0.045	0.043	0.050	0.042

Chapter 4

An Application on Time Series Models

4.1 Introduction

In this chapter we extend the approximations for the distribution of scan statistics, that have been derived in Wang and Glaz (2013) for iid normal observations, to time series models. These approximations will be evaluated for selective ARMA models with a Gaussian white noise component in Section 4.2. In Section 4.3, we present a multiple window scan statistic for detecting a local change in the mean of Gaussian white noise component of the underlying time series models. Based on Genz and Bretz (2009) algorithms for the multivariate normal distribution, we present algorithms for implementing this multiple window scan statistic. Moreover, for a specified model of a local change, we compare its power with that of several fixed window scan statistics. In Section 4.4, for selected values of parameters, numerical results are presented to evaluate the performance of proposed approximations and the effectiveness of the multiple window scan

statistic to detect a local change in the mean. We illustrate the performance of the proposed multiple window scan statistic on an observed data set. Concluding remarks are given in Section 4.5.

4.2 Approximations for $G(M)$

Let $X_1, ..., X_M$ be a sequence of observations from an autoregressive moving average (ARMA) time series model with a Gaussian white noise component with mean μ and variance σ^2. We assume that μ and variance σ^2 are known, or have been effectively estimated from a large data set. Let $Y_{i,j} = \sum_{t=i}^{j} X_t$, $1 \leq i < j \leq M$, denote a moving sum of length $j - i + 1$. For integers $2 \leq m < M$, where m is the size of the scanning window and M is the specified range to be scanned, define a *scan statistic*

$$S_{m,M} = \max_{m \leq j \leq M} \{Y_{j-m+1,j}\}. \tag{4.1}$$

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

$$G_{m,s}(M) = P(Y_{1,m} \leq s, Y_{2,m+1} \leq s, ..., Y_{M-m+1,M} \leq s).$$

The distribution of $S_{m,M}$ is given by

$$P(S_{m,M} \leq s) = G_{m,s}(M).$$

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.

The following approximation for $G(M)$ have been derived in Glaz et al. (2012):

$$G(M) \approx G(3m) \left[\frac{G(3m)}{G(2m)} \right]^{K-3}, \; K = [M/m]. \tag{4.2}$$

Therefore, to approximate the distribution of the scan statistic in (4.1), we need to evaluate accurately $G(2m)$ and $G(3m)$. Based on the assumption of Gaussian white noise component associated with an ARMA model, a joint multivariate normal distribution can be derived for time series observations.

We now proceed to discuss the representation of the covariance matrices for the specific ARMA models, for which the fixed window and multiple window scan statistics are investigated. Then, by utilizing an algorithm developed by Genz and Bretz (2009) for multivariate normal probabilities, we evaluate approximations for the distribution of fixed window scan statistic for time series observations. In Section 4.3, a multiple window scan statistic is developed based on approximation (4.2) for the distribution of a fixed window scan statistic in (4.1).

4.2.1 MA Models

Let $X_1, ..., X_M$ be a sequence of observations from an MA(1) model, $X_t = \omega_t + \theta\omega_{t-1}$, where ω_t is a Gaussian white noise with mean $\mu = 0$ and known variance σ^2. Without loss

of generality, we will assume $\sigma^2 = 1$. It is well known that the $X_t's$ have a multivariate normal distribution. Therefore, the moving sums $\{Y_{j-m+1,j}; m \leq j \leq M\}$ follow a multivariate normal distribution with a zero mean vector μ and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, where $\sigma_{i,j} = cov(Y_{i,i+m-1}, Y_{j,j+m-1})$. A routine derivation yields the following covariance matrix:

$$\sigma_{i,j} = \begin{cases} (j+m-i)(1+\theta^2) + 2\theta(j+m-i), & \text{if } i-j < m \\ \theta, & \text{if } i-j = m \\ m(1+\theta^2) + 2\theta(m-1), & \text{if } i = j \\ 0, & otherwise. \end{cases}$$

Given the mean vector and covariance matrix associated with the MA(1) model, one can utilize the algorithm developed by Genz and Bretz (2009) to approximate the distribution of $G(M)$ for a fixed window scan statistic.

Haiman and Preda (2013) introduced a different approximation for the distribution of fixed window scan statistics based on 1-dependent stationary sequences, which can be effectively applied for the MA(1) model. It follows from Haiman and Preda (2013), Theorem 1, that

$$G(M) \approx \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^K}, \; K = (M+1)/m - 1, \tag{4.3}$$

with an error bound:

$$Error = 3.3K(1-q_1)^2 + 2K * 1.96\sqrt{\frac{q_1(1-q_1)}{I}}, \tag{4.4}$$

where I, q_1 and q_2 are given in Haiman and Preda (2013), Section 3. In section 4.4, table 1, we present numerical results for approximations in (4.2) and (4.3).

For an MA(2) model, $X_t = \omega_t + \theta_1\omega_{t-1} + \theta_2\omega_{t-2}$, where ω_t is the Gaussian white noise with mean $\mu = 0$ and variance $\sigma^2 = 1$. Since $X_t's$ follow a multivariate normal distribution, with the following autocovariance function(ACF),

$$\gamma_h = \begin{cases} \sum_{j=0}^{2-h} \theta_j\theta_{j+h}, & \text{if } 0 \le h \le 2 \\ 0, & \text{if } h > 2. \end{cases}$$

It follows that $\{Y_{i-m+1,i}; m \le i \le M\}$ have a multivariate normal distribution with a mean vector of zeros and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, which can be derived in a similar manner as in the MA(1) case. For simplicity, we omit the presentation of its formulae here. In section 4.4, table 2, we evaluate the approximation in (4.2). Note that the approximation discussed in Haiman and Preda (2013) is valid only for the MA(1) model.

4.2.2 AR Models

Let $X_1, ..., X_M$ be a sequence of observations from an AR(1) process, $X_t = \theta X_{t-1} + \omega_t$, where ω_t is a Gaussian white noise with mean $\mu = 0$ and variance $\sigma^2 = 1$. Since $X_t's$

follow a multivariate normal distribution, $\{Y_{i-m+1,i}; m \leq i \leq M\}$ have a multivariate normal distribution with zero mean vector and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, where $\sigma_{i,j} = cov(Y_{i,i+m-1}, Y_{j,j+m-1})$. A routine derivation, yields the following covariance matrix:

$$\sigma_{i,j} = \begin{cases} \frac{\theta}{(1-\theta)^4}(1-\theta^{j+m-i})(1-\theta^{i-j}) + \frac{\theta^{j+m-i+1}}{(1-\theta)^4}(1-\theta^{i-j})^2 & \\ +\frac{j+m-i}{(1-\theta)^2} + \frac{2\theta}{(1-\theta)^3}[j+m-1-i-\frac{\theta}{1-\theta}(1-\theta^{j+m-1-i})] & \\ +\frac{\theta}{(1-\theta)^4}(1-\theta^{i-j})(1-\theta^{j+m-i}), & \text{if } i-j<m \\ \frac{1}{1-\theta^2}\{m+\frac{2\theta}{1-\theta}[m-1-\frac{\theta}{1-\theta}(1-\theta^{m-1})]\}, & \text{if } i=j \\ \theta^{i-j-m+1}\frac{(1-\theta^m)^2}{(1-\theta)^2}, & \text{otherwise.} \end{cases}$$

Given the mean vector and covariance matrix, we can utilize the algorithm developed by Genz and Bretz (2009) to approximate the distribution $G(M)$ for a fixed window scan statistic.

For an AR(2) model, $X_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + \omega_t$, where ω_t is the Gaussian white noise with mean $\mu = 0$ and variance $\sigma^2 = 1$,the $X_t's$ follow a multivariate normal distribution with the following ACF:

$$\gamma_h = \begin{cases} \frac{1-\theta_2}{1-\theta_2-\theta_1^2-\theta_2\theta_1^2-\theta_2^2+\theta_2^3}, & \text{if } h = 0 \\ \gamma_0 \frac{\theta_1}{1-\theta_2}, & \text{if } h = 1 \\ \gamma_0[\theta_1\gamma_{h-1} + \theta_2\gamma_{h-2}], & \text{if } h > 1. \end{cases}$$

The sequence of moving sums, $\{Y_{i-m+1,i}; m \leq i \leq M\}$, has a multivariate normal distribution with zero mean vector and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, which can be derived similarly as in the AR(1) process. For simplicity, the explicit form of the covariance matrix is omitted here.

4.2.3 ARMA Model

Let $X_1, ..., X_M$ be a sequence of observations from an ARMA(1,1) process, $X_t = \theta_1 X_{t-1} + \omega_t + \theta_2\omega_{t-1}$, where ω_t is the Gaussian white noise with mean $\mu = 0$ and variance $\sigma^2 = 1$. The $X_t's$ follow a multivariate normal distribution, and have the following ACF:

$$\gamma_h = \begin{cases} \frac{1+2\theta_1\theta_2+\theta_2^2}{1-\theta_1^2}, & \text{if } h = 0 \\ \frac{(1+\theta_1\theta_2)(\theta_1+\theta_2)}{1-\theta_1^2}\theta_1^{h-1}, & \text{if } h > 0. \end{cases}$$

Therefore, $\{Y_{i-m+1,i}; m \leq i \leq M\}$ have a multivariate normal distribution with a zero mean vector and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, which can be easily derived. For simplicity, its explicit form is omitted here.

4.3 A Multiple Window Scan Statistic

We now introduce a *multiple window scan statistic* for detecting a local change in the process mean of time series data, $X_t, 1 \leq t \leq M$, modeled by an ARMA model. We are interested in testing the null hypothesis H_0, that assumes the Gaussian white noise components in the time series model, ω_t, are iid normal random variables with mean $\mu = \mu_0 = 0$ and variance σ^2. For the alternative hypothesis, H_1, of a local change in μ , one often assumes that the following change has occurred in a segment of m consecutive observations,

$$R(t_0, m) = \{t_0, i_0 + 1, ..., t_0 + m - 1\},$$

where both the window size m, $2 \leq m \leq N/4$ and the location $t_0, 1 \leq t_0 \leq N - m + 1$, are unknown. Under H_1, for $t_0 \leq t \leq t_0 + m - 1$, ω_t has a normal distribution with mean $\mu = \mu_1$ and variance σ^2, where $\mu_1 > \mu_0$. For $t \notin R(t_0, m)$, ω_t's are distributed according to the distribution specified by the null hypothesis. When the length of the sliding window m is known, the generalized likelihood ratio test rejects the null hypothesis of randomness in favor of the local change alternative hypothesis, whenever S_m exceeds the value s, where s is determined from $P(S_m \geq s|H_0) = \alpha$, where α is a specified significance level of the testing procedure.

Since the size of the sliding window m is unknown, we propose to investigate the performance of a multiple window scan statistic based on a sequence of n fixed window scan statistics: $S_{m_1}, ..., S_{m_n}$, where $2 \leq m_1 < m_2 < .... < m_n \leq M/4$, where the lengths

of the n sliding windows, are chosen by the experimenter. For $1 \leq j \leq n$, let s_j be the observed value of S_{m_j} and $p_j = P(S_{m_j} \geq s_j \mid H_0)$ the associated p-value. To test H_0 vs H_1 we propose the following test statistic:

$$P_{\min} = \min\{p_j; 1 \leq j \leq n\}, \tag{4.5}$$

the *minimum P-value statistic*. Since the exact distribution for the P_{min} statistic is unknown, for a given significant level α, the critical value p_α,

$$P_{H_0}\left(P_{min} \leq p_\alpha \mid H_0\right) = \alpha,$$

has to be evaluated via simulation. In each run of the simulation, we generate M observations under the null hypothesis. Then we scan the whole region with multiple moving windows of sizes m_1, m_2, ... and m_n, and record the observed values of the fixed window scan statistics, $S_{m_1}, ..., S_{m_n}$, denoted by $s_1, s_2, ..., s_n$, respectively. At the next stage, a randomized quasi Monte-Carlo R algorithm by Genz and Bretz (2009) is employed to evaluate the observed p values $p_j = P(S_{m_j} \geq s_j \mid H_0)$, $1 \leq j \leq n$, by employing the approximation (4.2). The minimum of these p values is recorded and then the whole process is repeated N times. p_α will be calculated as the $\alpha * 100$ percentile of the simulated distribution of $P_{\min}$ statistic.

We now describe an algorithm to evaluate the performance of the P_{min} statistic to

detect a local change in the mean of a Gaussian white noise component in time series data and compare its power with that of fixed window scan statistics. In this algorithm we have specified the number of moving windows and their individual lengths. For the alternative hypotheses we specified the locally increased mean $\mu_1 > 0$, the location and the window length where the change in the mean process has occurred. For a specified significance level α, we have evaluated the power under each alternative hypothesis using the following steps:

1. Generate M observations from a specified ARMA time series model with a Gaussian white noise component with $\mu_0 = 0$ and $\sigma^2 = 1$. In a specified location t_0 and window of length m replace the observations obtained in step 1 with observations from an ARMA time series model that has Gaussian white noise component with mean $\mu_1 > 0$ and $\sigma^2 = 1$.

2. Scan the whole region with selected window sizes, and let $s_1, ..., s_n$ be the observed values of the fixed window scan statistics $S_{m_1}, ..., S_{m_n}$, respectively.

3. For a fixed window of length m_j and a specified significance level α, evaluate $p_j = 1 - P(S_{m_j} \leq s_j)$, and reject H_0 if $p_j < \alpha$.

4. For the multiple window scan statistic based on $P_{min} = min\{p_j; 1 \leq j \leq n\}$, reject H_0 if $P_{min} < p_\alpha$.

5. Repeat steps 1-4 T times and count how many times out of T, we have rejected

H_0 with both the fixed and multiple window scan statistics.

4.4 Numerical Results

In this Section, numerical results are presented to evaluate, for selected values of the parameters, the accuracy of approximations for the distribution of fixed window scan statistics for selected ARMA time series models, and to compare their power with the multiple window scan statistic, using a simulation study proposed in Section 4.3. Moreover, the use of a multiple window scan statistic in detecting a local change in Gaussian white noise component is illustrated for a Series D data set in Box and Jenkins (1976), page 529.

In Tables 4.1−4.5, for selected M, m, θ and threshold t, approximations are evaluated for tail probabilities $1-G(M)$, defined in Section 4.2, for time series observations with a Gaussian white noise component with mean 0 and variance 1. These numerical results are obtained from the R algorithm for the multivariate normal distribution in Genz and Bretz (2009). $APPRX1$, $APPRX2$ and $ErrorBound$ are evaluated via the approximations for $1-G(M)$ in (4.2), (4.3) and (4.4) respectively. To evaluate the approximation and error bound in Haiman and Preda (2013), 10^6 simulation runs have been used to estimate q_1 and q_2.

In Tables 4.6 − 4.11, numerical results are presented to evaluate the accuracy of achieving a specified probability of Type I error for the multiple window scan statistic,

and to compare its power with fixed window scan statistics. For a significance level $\alpha = 0.05$, the critical values p_α are evaluated via simulations with $10,000$ trials using the algorithms in Section 4.3. In Tables $4.6 - 4.10$, for selected values of parameters, we used $1,000$ replications to simulate the power of P_{min} and compare it with the power of fixed window scan statistics. The mean of Gaussian white noise under the null hypothesis μ_0 is 0, and $\mu_1 > 0$ within a consecutive sequence of m observations, under the alternative hypothesis. The power of P_{min} is compared to that of the fixed window scan statistics S_{m_j} where the length of the scanning window is m_j. Observations are generated under different time series models using a Monte Carlo simulation. In Table 4.11, we evaluate the performance of the multiple scan statistic, introduced in Section 4.3, in detecting a local change in the mean of Gaussian white noise, for the Series D data set in Box and Jenkins (1976), page 529. This data set consists of 310 hourly uncontrolled viscosity readings of a chemical process. To model this data set an AR(1) has been used in Box and Jenkins (1976), page 529, with estimated parameters $\theta = 0.87$ and $\sigma^2 = 0.09$. To evaluate the performance of the multiple window scan statistic, we have introduced a change in Gaussian white noise component at a random location. We employed steps $2 - 5$ in the algorithm outlined in Section 4.3, to perform a power study.

Based on the numerical results presented in Tables $4.1 - 4.5$, one can conclude that both approximations from (4.2) and (4.3) are quite accurate. The error bound for a MA(1) model will decrease as we increase the number of replications used to evaluate q_1 and q_2. The accuracy of both approximations is verified by a simulation study with

10^4 replications. By comparing the numerical results for power calculations in Tables $4.6 - 4.10$, fixed window scan statistics with correctly specified window size turned out to be most powerful in detecting a local change in Gaussian white noise mean. The multiple window scan statistic was slightly less powerful in that case, but outperformed the fixed window scan statistics with an incorrectly specified window size where a change in mean has occurred. In Table 4.11, we observed similar results, and the discrepancy could have resulted from the model lack of fit.

4.5 Conclusion

In this chapter, approximations for the distribution of fixed window scan statistics for observations generated by ARMA time series models have been discussed. Based on the numerical results these approximations appear to be quite accurate. A multiple window scan statistic has been introduced along with an algorithm for its implementation. The numerical results presented in Section 4.4, have been effectively evaluated via an R algorithm for the multivariate normal distribution in Genz and Bretz (2009). For detecting a local change in the mean of the observed time series data, generated by a local change in the Gaussian white noise component, it is evident that a multiple window scan statistic outperforms fixed window scan statistics, when the size of the region where a change in the mean has occurred is unknown.

Table 4.1: $P(S_{m,M} \geqslant t)$ for MA(1) process, M=1500, m=20, θ=0.1

t	18	19	20	21	22	23
APPRX 1	0.0659	.0294	.0192	.0060	.0034	.0007
APPRX 2	.0592	.0282	.0128	.0053	.0021	.0013
Error Bound	.0091	.0061	.0043	.0026	.0020	.0011
Simulation	.0622	.0288	.0135	.0055	.0021	.0007

Table 4.2: $P(S_{m,M} \geqslant t)$ for MA(2) process, M=500, m=10, $\theta_1 = \theta_2$=0.5

t	21	22	23	24	25	26
APPRX 1	.0500	.0424	.0112	.0030	.0015	.0029
BruteForce	.049	.0296	.0143	.0074	.0032	.0017

Table 4.3: $P(S_{m,M} \geqslant t)$ for AR(1) process, M=1500, m=20, θ=0.1

t	18	19	20	21	22	23
APPRX 1	.0695	.0347	.0168	.0050	.0036	.0028
BruteForce	.0686	.032	.016	.0063	.0032	.0018

Table 4.4: $P(S_{m,M} \geqslant t)$ for AR(2) process, M=500, m=10, $\theta_1 = \theta_2$=0.2

t	16	17	18	19	20	21
APPRX 1	.0734	.0496	.0238	.0087	.0006	.0036
BruteForce	.0945	.0489	.0245	.0105	.0041	.0019

Table 4.5: $P(S_{m,M} \geqslant t)$ for ARMA(1,1) process, M=1500, m=20, $\theta_1 = \theta_2$=0.1

t	19	20	21	22	23	24
APPRX 1	.1007	.0559	.0231	.0129	.0093	.0021
BruteForce	.1001	.0531	.0293	.0130	.0053	.0022

Table 4.6: one dimensional MA(1) variables, $\mu_0 = 0$, $M = 1500$, $\theta = 0.1$

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.5	0.065	0.070	0.073	0.055	0.059	0.051
	1	0.262	0.204	0.310	0.168	0.119	0.089
	1.5	0.797	0.655	0.848	0.563	0.386	0.279
15	0.5	0.088	0.071	0.091	0.097	0.079	0.074
	1	0.549	0.306	0.510	0.582	0.404	0.294
	1.5	0.965	0.809	0.947	0.972	0.914	0.823
20	0.5	0.130	0.078	0.098	0.123	0.136	0.117
	1	0.760	0.330	0.615	0.749	0.801	0.670
	1.5	0.998	0.889	0.990	0.998	1.000	0.993
Type I Error		0.051	0.058	0.066	0.058	0.039	0.054

Table 4.7: one dimensional MA(2) variables, $\mu_0 = 0$, $M = 500$,$\theta_1 = \theta_2 = 0.5$

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.5	0.080	0.079	0.093	0.067	0.070	0.063
	1	0.390	0.328	0.470	0.270	0.199	0.142
	1.5	0.900	0.816	0.933	0.747	0.567	0.434
15	0.5	0.116	0.098	0.124	0.141	0.105	0.084
	1	0.673	0.440	0.631	0.753	0.554	0.419
	1.5	0.994	0.938	0.984	0.997	0.975	0.928
20	0.5	0.194	0.106	0.160	0.193	0.209	0.157
	1	0.871	0.551	0.784	0.866	0.906	0.810
	1.5	0.998	0.976	0.996	0.998	0.998	0.997
Type I Error		0.047	0.078	0.042	0.047	0.054	0.048

Table 4.8: one dimensional AR(1) variables, $\mu_0 = 0$, $M = 1500$,$\theta = 0.1$

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.5	0.076	0.075	0.063	0.056	0.062	0.050
	1	0.292	0.211	0.337	0.172	0.124	0.101
	1.5	0.797	0.649	0.841	0.554	0.388	0.285
15	0.5	0.103	0.072	0.087	0.093	0.085	0.074
	1	0.532	0.273	0.494	0.594	0.407	0.297
	1.5	0.973	0.810	0.960	0.989	0.915	0.800
20	0.5	0.115	0.068	0.091	0.113	0.120	0.100
	1	0.762	0.378	0.615	0.759	0.818	0.683
	1.5	1.000	0.905	0.992	1.000	1.000	0.992
Type I Error		0.050	0.037	0.052	0.052	0.051	0.052

Table 4.9: one dimensional AR(2) variables, $\mu_0 = 0$, $M = 500$,$\theta_1 = \theta_2 = 0.2$

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.5	0.093	0.095	0.103	0.072	0.074	0.073
	1	0.486	0.423	0.529	0.294	0.191	0.150
	1.5	0.935	0.884	0.957	0.760	0.578	0.458
15	0.5	0.150	0.126	0.140	0.154	0.108	0.092
	1	0.724	0.536	0.690	0.778	0.584	0.462
	1.5	0.996	0.968	0.986	0.998	0.982	0.912
20	0.5	0.200	0.132	0.166	0.176	0.210	0.160
	1	0.892	0.688	0.818	0.880	0.922	0.822
	1.5	1.000	0.994	1.000	1.000	1.000	1.000
Type I Error		0.051	0.054	0.036	0.051	0.040	0.060

Table 4.10: one dimensional ARMA(1,1) variables, $\mu_0 = 0$, $M = 1500$,$\theta_1 = \theta_2 = 0.1$

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.5	0.068	0.074	0.072	0.068	0.055	0.056
	1	0.255	0.213	0.306	0.167	0.116	0.099
	1.5	0.825	0.666	0.870	0.630	0.420	0.298
15	0.5	0.095	0.074	0.074	0.104	0.064	0.060
	1	0.525	0.291	0.487	0.593	0.422	0.307
	1.5	0.965	0.828	0.951	0.985	0.923	0.824
20	0.5	0.123	0.062	0.107	0.122	0.136	0.107
	1	0.749	0.364	0.622	0.744	0.811	0.689
	1.5	0.998	0.892	0.990	0.997	0.999	0.993
Type I Error		0.049	0.054	0.054	0.054	0.049	0.046

Table 4.11: real data AR(1) process, $\mu_0 = 0$, $M = 310$,$\theta = 0.87$, $\sigma^2 = 0.09$

m	μ_1	P_{min}	S_5	S_{10}	S_{15}	S_{20}	S_{25}
10	0.15	0.142	0.170	0.260	0.035	0	0
	0.2	0.584	0.588	0.628	0.330	0.100	0
	0.25	0.703	0.731	0.710	0.641	0.403	0.262
15	0.15	0.394	0.234	0.379	0.473	0.226	0.161
	0.2	0.704	0.725	0.700	0.664	0.622	0.520
	0.25	0.841	0.845	0.834	0.864	0.756	0.731
20	0.15	0.625	0.324	0.499	0.574	0.617	0.500
	0.2	0.778	0.787	0.774	0.750	0.753	0.737
	0.25	0.932	0.889	0.914	0.910	0.942	0.805
Type I Error		0.051	0.040	0.054	0.049	0.059	0.040

Bibliography

Alm, S. (1997). On the ditribution of scan statistics of two-dimensional poisson processes. *Advances in Applied Probability*, 29:1–28.

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

Alm, S. (1999). Approximations of the distributions of scan statistics of poisson processes. In *Scan Statistics and Applications*, pages 113–139. Birkhäuser Boston.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56.

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

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

Boutsikas, M. and Koutras, M. (2000). Reliability approximations for markov chain imbeddable systems. *Methododlogy and Computing in Applied Probability*, 2(4):393–411.

Box, G. and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics.

Chan, H. (2009). Maxima of moving sums in a poisson random field. *Advances in Applied Probability*, 41(3):647–663.

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

Chen, J. and Glaz, J. (1999). Approximation for the distribution of the moments of discrete scan statistics. In *Scan Statistics and Application*, pages 27–66. Birkhäuser Boston.

Chen, J. and Glaz, J. (2002). Approximations for a conditional two-dimensional scan statistic. *Statistics and Probability letters*, 58(3):287–296.

Chen, J. and Glaz, J. (2009). Approximations for two-dimensional variable window scan statistic. In *Scan StatisticsMethods and Applications*, pages 109–128. Birkhäuser Boston.

Chu, C., Hornik, K., and Kaun, C. (1995). Mosum tests for parameter constancy. *Biometrika*, 82:603–617.

Cressie, N. (1991). *Statistics for Spatial Data*. Wiley, New York.

Darling, R. and Waterman, M. (1986). Extreme value distribution for the largest cube in a random lattice. *SIAM Journal on Applied Mathematics*, 46:118–132.

Dominici, F., McDermott, A., Zeger, S., and Samet, J. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, 156(3):193–203.

Friston, K., Williams, S., Howard, R., Frackowiak, R., and Turner, R. (2011). Movement-related effects in fmri time-series. *Magnetic Resonance in Medicine*, 35(3):346–355.

Fu, J. and Koutras, M. (1994). Distribution theory of runs: A markov chain approach. *Journal of the American Statistical Association*, 89:1050–1058.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and T Probabilities*. Springer.

Glaz, J. (1996). Discrete scan statistics with applications to minefields detection. In *Preceedings of Conference SPIE*, pages 420–429, Orlando, Florida.

Glaz, J. and Johnson, B. (1988). Boundary crossing for moving sums. *Journal of Applied Probability*, 25:81–88.

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

Glaz, J., Naus, J., and Wallenstein, S. (2001). *Scan Statistics*. Springer.

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

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

Guerriero, M., Willett, P., and Glaz, J. (2009). Distributed target detection in sensor networks using scan statistics. *IEEE Transactions on Signal Processing*, 57(7):2629–2639.

Haiman, G. (1999). First passage time for some stationary process. *Stochastic Processes and their Applications*, 80:231–248.

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

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

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

Haiman, G. and Preda, C. (2013). One dimensional scan statistics generated by some dependent stationary sequences. *Statistics and Probability Letters*, 83:1457–1463.

Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for suspectibility genes. In *Proceedings of the National Academy of Sciences USA*, volume 97, pages 9615–9617.

Holst, L. and Janson, S. (1990). Poisson approximations using the stein-chen method and coupling: number of exceedances of gaussian random variables. *Annals of Probability*, 18:713–723.

Hunter, D. (1976). An upper bound for the probability of a union. *Journal of Applied Probability*, 13:597–603.

J.D. Esary, F. P. and Walkup, D. (1967). Association of random variables. *Annals of Mathematical Statistics*, 38:1466–1474.

Koen, C. (1991). A computer program package for the statistical analysis of spatial point processes in a square. *Biometric Journal*, 33:493–503.

Kulldorff, M. (1997). A spatial scan statistic. *Communication in StatisticsTheory and Methods*, 26:1481–1496.

Kwerel, M. (1975). Most stringent bounds on the probability of the union and intersection of m events. *Journal of Applied Probability*, 12:612–619.

Lai, T. (1974). Control charts based on weighted sums. *Annals of Statistics*, 2:134–147.

Malinowski, J. and Preuss, W. (1995). Reliability of circular consecutively-connected systems with multistate components. *IEEE Transactions on Reliability*, 44(3):532–534.

Naus, J. (1982). Approximations for distributions of scan statistics. *Journal of the American Statistical Association*, 77:177–183.

Naus, J. and Sheng, K. (1996). Screening for unusual matched segments in multiple protein sequences. *Communications in Statistics*, 25:937–952.

Panayirci, E. and Dubes, R. (1983). A test for multidimensional clustering tendency. *Pattern Recognitionl*, 16:433–444.

Patil, G., Joshi, S., Myers, W., and Koli, R. (2009). Uls scan statistics for hotspot detection with continuous gamma response. In *Scan Statistics*, pages 251–270. Birkhäuser Boston.

Pfaltz, J. (1983). Convex clusters in discrete m-dimensional space. *Journal of Computation*, 12:746–750.

Rosenfeld, A. (1978). Clusters in digital pictures. *Information and Control*, 39:19–34.

Saperstein, B. (1976). The analysis of attribute moving averages: Mil-std-105d reduced inspection plans. In *Proceedings of Sixth Conference Stochastic Processes and Applications*. Tel-Aviv University.

Song, X., Willett, P., Glaz, J., and Zhou, S. (2012). Distributed detection with a scan statistic: Global to local inference. In *Sensor Array and Multichannel Signal Processing Workshop(SAM), 2012 IEEE 7th*, pages 485–488.

Waldman, K. (1986). Bounds to the distribution of the run length in general quality-control schemes. *Staistische Hefte*, 27:37–56.

Wang, X. and Glaz, J. (2013). Variable window scan statistics for normal data. *Communications in Statistics - Theory and Methods, to appear.*

Webster, P., Holland, G., Curry, J., and Chang, H. (2005). Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, 309(5742):1844–1846.

Westlund, A. (1984). Sequential moving sums of squares of ols residuals in parameter stability testing. *Quality and Quantity*, 18:261–273.

Xia, Z., Guo, P., and Zhao, W. (2009). Monitoring structural changes in generalized linear models. *Communications in Statistics: Theory and Methods*, 38:1927–1947.

Zhang, Z. and Glaz, J. (2008). Bayesian variable window scan statistics. *Journal of Statistical Planning and Inference*, 138(11):3561–3567.