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

Joseph Glaz

University of Connecticut

September 2017

Outline of the Presentation

- Scan statistics for local change in variance for normal data.
  - Introduction
  - Approximations for one dimensional data: population variance is known
  - Scan statistics for one dimensional data when the variance known.
  - Multiple window scan for one dimensional data when the variance is known.
  - Scan statistics for two dimensional data
- Summary and future work.

Introduction: Early References on Clustering of Events

- Berg, W. (1945). Aggregates in one-and-two-dimensional random distributions. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **36**, 337-346.
- Dinneen, G. P. and Reed, I. S. (1956). An analysis of signal detection and location by digital methods. *IRE Trans. Information Theory*, **IT-2**, 29-39.
- Domb, C. (1950). Some probability distributions connected with recording apparatus II. *Proceedings Cambridge Phil. Soc.*, **46**, 429-435.
- Mack, C. (1948). An exact formula for $Q_k(n)$, the probable number of k- aggregates in a random distribution of n points. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **39**, 778-790.
- Silberstein, L. (1945). The probable number of aggregates in random distributions of points. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **36**, 319-336.

Introduction: Early Theoretical Advances on Scan Statistics

- Naus, J. I. (1963). *Clustering of Random Points on the Line and Plane*. Ph.D. Thesis, Harvard University, Cambridge, MA.
- Naus, J. I. (1965a). The distribution of the size of the maximum cluster of points on a line. *J. Amer. Stat. Assoc.* **60**, 532-538.
- Naus, J. I. (1965b). Power comparison of two tests of non-random clustering. *Technometrics* **8**, 493-517.
- Naus, J. I. (1965b). Power comparison of two tests of non-random clustering. *Technometrics* **8**, 493-517.
- Barton, D. E. and Mallows, C. L. (1965). Some aspects of the random sequence. *Annals of Mathematical Statistics* **36**, 236-260.
- Karlin, S. and McGregor, G. (1959). Coincidence probabilities. *Pacific Journal of Mathematics* **9**, 1141-1164.

Introduction: Early Theoretical Advances on Scan Statistics

- Wallenstein, S. and Naus, J. (1973). Probabilities of kth nearest neighbor problem on the line. *Annals of Probability* **1**, 188-190.
- Wallenstein, S. and Naus, J. (1974). Probabilities for the size of the largest clusters and smallest intervals. *J. American Statist. Assoc.* **69**, 690-697.
- Cressie, N. (1977). On some properties of the scan statistic on the circle and the line. *J. Applied Probability* **14**, 272-283.
- Glaz, J. (1979). Expected waiting time for the visual response. *Biological Cybernetics* **35**, 39-41.
- Glaz, J. and Naus J. (1979). Multiple coverage of the line. *Annals of Probability* **7**, 900-906.
- Huntington, R. J. and Naus, J. I. (1975). A simpler expression for the Kth nearest neighbor coincidence probabilities. *Annals of Probability* **3**, 894-896.

Introduction: One dimensional data

- Let $X_1, \ldots, X_M$ be a sequence of iid normal observations with mean μ and variance σ^2, where M is the specified range of the monitoring process. We are interested in detecting a local upward shift in variance.
- Let $2 \leq m \leq M/4$, be the size of the sliding window of a segment of m consecutive observations. We are interested in testing the following hypotheses:
- H_0: $X_i, 1 \leq i \leq M$, are iid normal random variables with mean μ and variance σ_0^2, vs. H_a: $X_i, 1 \leq i \leq M$, are independent normal random variables with mean μ, the $X_i' s$ have variance $\sigma_1^2 > \sigma_0^2$,
- for $i \in R(a, m) = \{a, a+1, ..., a+m-1\}$, where $1 \leq a \leq M - m + 1$ is unknown, and variance σ_0^2 for $i \notin R(a, m)$. The restriction $m \leq M/4$ is used to emphasize the interest in detecting a local change in variance.
- In the above hypotheses one can always assume that $\mu = 0$. If $\mu \neq 0$, one can replace the $X_i' s$ with the sequence of recurrent residuals:

Introduction: One dimensional data

- $$W_i = \frac{(i-1)X_i - \sum_{j=1}^{i-1} X_j}{\sqrt{i(i-1)}}, 2 \le i \le M,$$
  which are iid normal random variables with mean 0 and variance σ_0^2, under the null hypothesis (Bauer 1978).
- When σ_0^2 is known, without loss of generality one can assume $\sigma_0^2 = 1$.
- A *scan statistic* for detecting a local change in variance, is defined by:

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

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

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

Under H_0, the random variables $Y_{r,m}, 1 \le r \le M - m + 1$, are m-dependent and have a joint multivariate chi-square distribution and marginal chi-square distributions with m degrees of freedom.

The joint covariance matrix is given by: $\Sigma = \{\sigma_{i,j}\}$, where: $\sigma_{i,i} = 2m$, for $1 \leq i \leq m, \sigma_{i,j} = 0$, for $|j - i| \geq m$ and $\sigma_{i,j} = 2(m - k)$, for $|j - i| = k$, $1 \leq k \leq m - 1$. For $2 \leq m \leq M/4$ and $-\infty < t < \infty$, let

$$G_{m,t}(M) = P(S_{m,M} < t) = P(Y_{1,m} < t, Y_{2,m} < t, \ldots, Y_{M-m+1,m} < t), \quad (3)$$

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

$$P(S_{m,M} \geq t) = 1 - G_{m,t}(M). \quad (4)$$

Introduction: One dimensional data

- For our hypotheses testing problem, when the window size m is known, the generalized likelihood ratio test rejects the null hypothesis, in favor of the local change alternative hypothesis H_a,
- whenever $S_{m,M}$ exceeds a threshold value t, where t is determined by $P(S_{m,M} \geq t|H_0) = \alpha$, α being the specified significance level.
- Hence, to implement our testing procedure we need to evaluate $G(M)$.
- Unlike the case of detecting a local change in the mean level for the normal data, where extensive theoretical results and R algorithms for computing multivariate normal and t distributions are readily available
- (Genz 2009 and Wang and Glaz 2014), for the problem at hand there are no algorithms to evaluate $G(M)$.
- Due to complexity of the dependence structure of the multivariate chi-square distribution for $Y_{r,m}, 1 \leq r \leq M - m + 1$, one has to evaluate $G(M)$ via Monte Carlo simulation.

Introduction: Two Dimensional Data

- For $1 \le i \le M_1$ and $1 \le j \le M_2$, let $\{X_{ij}\}$ be iid normal observations with mean μ and variance σ_0^2. We are interested in detecting an occurrence of a local change in variance, from σ_0^2 to σ_1^2, within a rectangular subregion of $m_1 \times m_2$ observations.
- For $k = 1, 2$, let $2 \le m_k \le M_k/4$ be the pre-specified size of a two dimensional sliding window. A fixed window *scan statistic* for detecting a local change in variance, is defined by:
- $$S_{m_1,m_2}(M_1, M_2) = max\{Y_{i_1,i_2}(m_1, m_2); 1 \le i_k \le M_k - m_k + 1, k = 1, 2\} \quad (5)$$

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

  $$Y_{i_1,i_2}(m_1, m_2) = \sum_{i=i_1}^{i_1+m_1-1} \sum_{i=i_2}^{i_2+m_2-1} X_{ij}^2 \quad (6)$$

  are the moving sums of squares in the $m_1 \times m_2$ rectangular grid of the observed data with south west location (i_1, i_2).

Introduction: Two Dimensional Data

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

$$G_{m,t}(M) = P(S_{m,m}(M,M) \leq t) = P(max\{Y_{i_1,i_2}(m_1,m_2); 1 \leq i_k \leq M_k \tag{7}$$

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

- Then,

$$P(S_{m,m}(M,M) > t) = 1 - G_{m,t}(M). \tag{8}$$

- We test the null hypothesis: H_0: $X_{ij}, 1 \leq i,j \leq M$, are iid. normal observations with mean μ and variance σ_0^2. The alternative hypothesis is: H_a: $X_{ij}, 1 \leq i,j \leq M$, are independent normal observations with mean μ, the $X_{ij}'s$ have variance $\sigma_1^2 > \sigma_0^2$, for $i,j \in R_{a_1,a_2}(m,m) = \{(i_1,i_2); a_k \leq i_1, i_2 \leq a_k + m + 1, k = 1,2\}$, where $1 \leq a_1, a_2 \leq M - m + 1$ are unknown coordinates of the southwest location of an $m \times m$ window, and variance σ_0^2 for $i,j \notin R_{a_1,a_2}(m,m)$.
- For our hypotheses testing problem, without loss of generality, one can always assume that $\mu = 0$ and $\sigma_0^2 = 1$.

Introduction: two dimensional data

- When the true window size m where a change in variance has occurred, is known, the generalized likelihood ratio test rejects our null hypothesis,
- in favor of the local change alternative hypothesis H_a, whenever $S_{m,m}$ exceeds a threshold value t, where t is determined
- by $P(S_{m,m} \geq t | H_0) = \alpha$, where α is the specified significance level.
- Hence, to implement our testing procedure we need to evaluate accurately $G(M)$, the joint distribution of the moving sum of squares.
- Under H_0, the random variables $\{Y_{i_1,i_2}(m_1, m_2); 1 \leq i_k \leq M_k - m_k + 1, k = 1, 2\}$, are m^2 -dependent and have a joint multivariate chi-square distribution and marginal chi-square distributions with m^2 degrees of freedom.
- To expedite the computations, two approximations by based on Wang and Glaz (2014) or Haiman (2006) can be used.

Approximations for G(m)

- We now present two approximations for $G(M)$. It follows from Glaz, Naus and Wang (2012), that:

$$G(M) = G(3m)\left[\frac{G(3m)}{G(2m)}\right]^{K-3}\frac{G(2m+v)}{G(2m)}, \tag{9}$$

  where $K \geq 3$, $m \geq 2$ and $0 \leq v \leq m-1$ are integers such that $M = Km + v$.
- The second approximation for $G(M)$ is based on Haiman (2007):

$$G(M) = \frac{2G(2m) - G(3m)}{[1 + G(2m) - G(3m) + 2(G(2m) - G(3m))^2]^{M/m-1}}, \tag{10}$$

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

$$3.3[1 - G(2m)]^2(M/m - 1), \tag{11}$$

  $M \geq 3m$, $1 - G(2m) \leq 0.025$ and
  $3.3M[1 - G(2m)]^2(M/m - 1) \leq 1$.

Approximations for G(m)

- The two approximations reduce significantly the computing time to evaluate $G(m)$, especially when M/m is large.

Approximations for G(m)

- The two approximations reduce significantly the computing time to evaluate $G(m)$, especially when M/m is large.
- Monte Carlo simulation is used to evaluate the accuracy of these two approximations.

Approximations for G(m)

- The two approximations reduce significantly the computing time to evaluate $G(m)$, especially when M/m is large.
- Monte Carlo simulation is used to evaluate the accuracy of these two approximations.
- An effective algorithm has been developed in Zhao and Glaz (2016a) to search for the critical value that determines the rejection region.

Approximations for G(m)

- The two approximations reduce significantly the computing time to evaluate $G(m)$, especially when M/m is large.
- Monte Carlo simulation is used to evaluate the accuracy of these two approximations.
- An effective algorithm has been developed in Zhao and Glaz (2016a) to search for the critical value that determines the rejection region.
- Numerical examples.

Scan Statistics for One Dimensional Data - Variance Known

- The performance of a fixed window scan statistic is evaluated in Zhao and Glaz (2016a, Section 2).
- We now outline the steps for deriving a variable window scan statistic. For a local upward shift in variance, the generalized likelihood ratio test will reject H_0 in favor of H_a for large values of

$$\Lambda = \frac{\sup_{\theta \epsilon \Theta_1} \prod_{i=1}^{M} f_\theta(x_i)}{\sup_{\theta \epsilon \Theta_0} \prod_{i=1}^{M} f_\theta(x_i)}, \tag{12}$$

where $f_\theta(x_i)$ is the probability density of the ith observation in the scanned sequence $\{X_i\}$ and Θ_0 and Θ_1 are the parameter spaces for the null and alternative hypotheses, respectively.

- This generalized likelihood ratio statistic can be expressed as follows:

Scan Statistics for One Dimensional Data - Variance Known

- $$\begin{aligned}\Lambda &= \sup_{\Theta_1} \left(\frac{1}{\sigma_1}\right)^m exp\left(\frac{1}{2}\sum_{i=a}^{a+m-1} X_i^2 - \frac{1}{2\sigma_1^2}\sum_{i=a}^{a+m-1} X_i^2\right) \\ &= \sup_{\Theta_1} \left(\frac{1}{\sigma_1}\right)^m exp\left(\frac{1}{2}Y_{a,m} - \frac{1}{2\sigma_1^2}Y_{a,m}\right) \\ &= \sup_{a;m} \left(\frac{m}{Y_{a,m}}\right)^{m/2} exp\left(\frac{1}{2}Y_{a,m} - \frac{m}{2}\right), \end{aligned} \tag{13}$$

  where $Y_{a,m} = \sum_{i=a}^{a+m-1} X_i^2$.
- The last step follows from the fact that for fixed and but arbitrary a and m, constrained by parameter space Θ_1, the supremum is achieved at $\hat{\sigma}_1^2 = Y_{a,m}/m > 1$.

Scan Statistics for One Dimensional Data - Variance Known

- Let

$$L_m(Y_{a,m}) = \left(\frac{m}{Y_{a,m}}\right)^{m/2} exp\left(\frac{1}{2}Y_{a,m} - \frac{m}{2}\right). \quad (14)$$

- Regard $L_m(Y_{a,m})$ as a function of $Y = Y_{a,m}$, depending on a, for fixed but arbitrary m. This function is a convex function of Y and it is increasing in Y on Θ_1.
- Therefore, for fixed m, the supremum in (13) is achieved at the maximum value of Y. One can obtain a unique value of a that maximizes Y.
- It follows that, for a given sequence of observations, one can get the location and length of the window that maximizes $L_m(Y_{a,m})$.
- This maximum value of $L_m(Y_{a,m})$ is the value of our variable window scan statistic based on the generalized likelihood ratio principle.
- An algorithm in Zhao and Glaz (2016a) implements the search for the location and length of the window that maximizes $L_m(Y_{a,m})$.

Scan Statistics for One Dimensional Data - Variance Known

- If M is large, implementing a variable window scan statistics might be computationally intensive. We propose to investigate the performance of the following multiple window scan statistic, based on the *minimum P-value* method (Glaz and Zhang 2004 and Wang and Glaz 2014). Since the window length m is unknown, a sequence of n fixed window scan statistics $\{S_{m_1}, S_{m_2}, \ldots, S_{m_n}\}$ can be employed simultaneously, where $2 \leq m_1 < m_2 < \ldots < m_n \leq M/4$.
- The lengths of the n sliding windows are chosen in advance by the experimenter. For $1 \leq j \leq n$, let t_j be the observed value of S_{m_j} and $p_j = P(S_{m_j} > t_j | H_0)$ its associated p-value.
- To test H_0 vs. H_a, the minimum p-value statistic, P_{min}, is defined as follows:

$$P_{min} = \min\{p_j; 1 \leq j \leq n\}. \tag{15}$$

Scan Statistics for One Dimensional Data - Variance Known

- The null hypothesis is rejected if the observed value of P_{min} falls below a critical value corresponding to a specified significance level α.
- Since the exact distribution of the P_{min} statistic is unknown, for a given significant level α, the critical value p_α, $P_{H_0}(P_{min} < p_\alpha) = \alpha$, has to be evaluated by a Monte Carlo simulation.
- The following algorithm can be used to find the critical value p_α.
- An algorithm to implement the P_{min} statistic has been presented in Zhao and Glaz (2016a).
- Numerical results to compare the performance of the variable and multiple window scan statistics.

Scan Statistics for One Dimensional Data - Variance Unknown

- A Training sample approach is discussed in Zhao and Glaz (2016b).
- A second approach to eliminate the unknown parameter σ_0^2, when H_0 is true, is to condition on the sufficient statistic under H_0.
- When H_0 is true, the sufficient statistic for σ_0^2 under H_0 is:

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

- The distribution of the random vector $\{X_i, 1 \leq i \leq M\}$, given $R^2 = r^2$, is uniform on a sphere of radius r (Dempster 1969, Chap. 12). Moreover, the random vector

$$\{X_i^{**} = X_i / R; 1 \leq i \leq M\}, \quad (16)$$

where $R = \sqrt{\sum_{i=1}^{M} X_i^2}$, has a joint uniform distribution on the $(M-1)$ dimensional unit sphere.

Scan Statistics for One Dimensional Data - Variance Unknown

- Consequently, we can define a scan statistic for the sequence of observations $\{X_i^{**}; 1 \leq i \leq M\}$:

$$S_{m,M}^{**} = max\{Y_{r,m}^{**}; 1 \leq r \leq M - m + 1\}, \qquad (17)$$

- where

$$Y_{r,m}^{**} = \sum_{i=r}^{r+m-1} X_i^{**2} = \frac{\sum_{i=r}^{r+m-1} X_i^2}{R^2}; 1 \leq r \leq M - m + 1. \qquad (18)$$

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

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

where s is the observed value of $S_{m,M}^{**}$.

Scan Statistics for One Dimensional Data - Variance Unknown

- Under H_0, the distribution of $S^{**}_{m,M}$ does not depend on any unknown parameters.
- Hence, for a given significance level α we can find the critical value t such that $P(S^{**}_{m,M} \geq t | R^2 = r^2, H_0) = \alpha$.
- These computations will be implemented via a Monte Carlo simulation that generates N sequences of data of M iid $N(0,1)$ observations, and then dividing each observation by R.
- Numerical results to evaluate the performance of the fixed window scan statistic via the approach of conditioning on the sufficient statistic.

Scan Statistics for One Dimensional Data - Variance Unknown

- A third approach for our testing problem is a parametric bootstrap test.
- The first step in implementing it is to estimate σ_0^2, the unknown population variance under H_0, via the sample variance of the observed data: $\widehat{\sigma_0^2} = S_M^2$.
- Let $\widehat{F}_0$ denote the fitted null model based on this estimate of σ_0^2.
- The method of evaluation of the P-value for the fixed window scan statistic $S_{m,M}$, where s is its observed value, via:

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

  is referred to as a parametric bootstrap test.
- This P-value is evaluated via simulation, based on an algorithm in Zhao and Glaz (2016b).

Multiple and Variable Window Scan Statistics

- Based on numerical results for fixed window scan statistics, it is evident that the scan statistic conditional on the sufficient statistic for σ_0^2 is superior to the other two fixed window scan statistics.

Multiple and Variable Window Scan Statistics

- Based on numerical results for fixed window scan statistics, it is evident that the scan statistic conditional on the sufficient statistic for σ_0^2 is superior to the other two fixed window scan statistics.
- Hence, multiple and variable window scan statistics will be based on the method of conditioning on the sufficient statistic for σ_0^2

Multiple and Variable Window Scan Statistics

- Based on numerical results for fixed window scan statistics, it is evident that the scan statistic conditional on the sufficient statistic for σ_0^2 is superior to the other two fixed window scan statistics.
- Hence, multiple and variable window scan statistics will be based on the method of conditioning on the sufficient statistic for σ_0^2
- For $n \geq 2$, let $2 \leq m_1 < m_2 < \ldots < m_n \leq M/4$ be the n sliding windows. For the transformed data sequence $\{X_1^{**}, ..., X_M^{**}\}$, defined in (16), the corresponding fixed window scan statistics, $S_{m_1,M}^{**}, S_{m_2,M}^{**}, \ldots, S_{m_n,M}^{**}$, are given in equation (17).

Multiple and Variable Window Scan Statistics

- Based on numerical results for fixed window scan statistics, it is evident that the scan statistic conditional on the sufficient statistic for σ_0^2 is superior to the other two fixed window scan statistics.
- Hence, multiple and variable window scan statistics will be based on the method of conditioning on the sufficient statistic for σ_0^2
- For $n \geq 2$, let $2 \leq m_1 < m_2 < \ldots < m_n \leq M/4$ be the n sliding windows. For the transformed data sequence $\{X_1^{**}, ..., X_M^{**}\}$, defined in (16), the corresponding fixed window scan statistics, $S_{m_1,M}^{**}, S_{m_2,M}^{**}, \ldots, S_{m_n,M}^{**}$, are given in equation (17).
- For $1 \leq j \leq n$, let s_j,be the observed value of $S_{m_j,M}^{**}$ and $p_j = P(S_{m_j,M}^{**} > s_j | R^2 = r^2, H_0)$ its associated p-value, respectively.

Multiple and Variable Window Scan Statistics

- Based on numerical results for fixed window scan statistics, it is evident that the scan statistic conditional on the sufficient statistic for σ_0^2 is superior to the other two fixed window scan statistics.
- Hence, multiple and variable window scan statistics will be based on the method of conditioning on the sufficient statistic for σ_0^2
- For $n \geq 2$, let $2 \leq m_1 < m_2 < \ldots < m_n \leq M/4$ be the n sliding windows. For the transformed data sequence $\{X_1^{**}, ..., X_M^{**}\}$, defined in (16), the corresponding fixed window scan statistics, $S_{m_1,M}^{**}, S_{m_2,M}^{**}, \ldots, S_{m_n,M}^{**}$, are given in equation (17).
- For $1 \leq j \leq n$, let s_j,be the observed value of $S_{m_j,M}^{**}$ and $p_j = P(S_{m_j,M}^{**} > s_j | R^2 = r^2, H_0)$ its associated p-value, respectively.
- For testing H_0 vs. H_1, we employ the following minimum P-value statistic, denoted by P_{min}:

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

- P_{min} is referred to as a *conditional multiple window scan statistic*. Note that, in the context of multiple testing, P_{min} can be viewed as a bootstrap test statistic (Davison and Hinkley 1997, Sec. 4.4.3).
- Since the exact distribution of the P_{min} statistic is unknown, for a given significant level α, the critical value p_α :

$$P_{H_0}(P_{min} < p_\alpha) = \alpha, \tag{19}$$

  has to be computed by a Monte Carlo simulation.
- While employing P_{min} to test H_0 vs. H_1, one can obtain an estimate of the window size where a change in variance has occurred, $\hat{m}$, from the window size corresponding to the observed value of P_{min}. Moreover, one can estimate the starting location of the window with the change of variance, $\hat{i}_0$, via the location which maximizes the moving sum squares with the fixed window size $\hat{m}$.

- A related test statistic is a conditional generalized likelihood ratio test (GLRT), based on conditioning on the total sum of squares of the whole data sequence, $\sum_{k=1}^{M} X_k^2 = R^2$, and the sum of squares of the partial data, $\{X_{i_0}, \ldots, X_{i_0+m-1}\}$ corresponding to a specified alternative, $\sum_{k=i_0}^{i_0+m-1} X_k^2 = r^2$, where $3 \leq m \leq M/4$.
- Therefore, under H_0, $(X_1, X_2, \ldots, X_M)$, conditional on R, has a joint the uniform distribution on the $(M-1)$ sphere with radius R.
- Under H_1, conditional on R and r, $(X_1, \ldots, X_{i_0-1}, X_{i_0+m}, \ldots, X_M)$ jointly follow a uniform distribution on the $(M-m-1)$ sphere with radius $\sqrt{R^2 - r^2}$ and are independent of $(X_{i_0}, \ldots, X_{i_0+m-1})$, where the latter jointly follow the uniform distribution on $(m-1)$ sphere with radius r.

Multiple and Variable Window Scan Statistics

- Hence, for the problem at hand, the conditional GLRT is given by:

$$\Lambda = \frac{sup_{\Theta_1} f(x_1, \ldots, x_M \mid R, r)}{sup_{\Theta_0} f(x_1, \ldots, x_M \mid R, r)}$$
$$= \frac{sup_{\Theta_1}\left\{\frac{1}{SS_{m-1}(r)} \times \frac{1}{SS_{M-m-1}(\sqrt{R^2-r^2})}\right\}}{sup_{\Theta_0}\left\{\frac{1}{SS_{M-1}(R)}\right\}}$$
$$= \frac{sup_{\Theta_1}\left\{\frac{1}{m\pi^{m/2}r^{m-1}/\Gamma(m/2+1)(M-m)\pi^{(M-m)/2}(R^2-r^2)^{(M-m-1)/2}/\Gamma(M/2-m/2}\right.}{sup_{\Theta_0}\left\{\frac{1}{M\pi^{M/2}R^{M-1}/\Gamma(M/2+1)}\right\}} \quad (20)$$

- where $f(x_1, \ldots, x_M \mid R, r)$ is the joint density of $X_1, \ldots, X_M$ conditional on R and r under respective hypotheses; $SS_N(K) = N\pi^{N/2}K^{N-1}/\Gamma(N/2+1)$ gives the surface area of the $(N-1)$ dim sphere with radius K.

Multiple and Variable Window Scan Statistics

- After routine derivations, it follows from equation (20), that

$$\begin{aligned}\Lambda =&\Lambda(m, i_0 \mid R, r)\\ \propto& sup_{\Theta_1} \frac{B(m/2,(M-m)/2)}{(r^2/R^2)^{(m-1)/2}(1-r^2/R^2)^{(M-m-1)/2}}\\ \propto& sup_{\{m,i_0\}} \frac{B(m/2,(M-m)/2)}{(Y^{**}_{i_0,m})^{(m-1)/2}(1-Y^{**}_{i_0,m})^{(M-m-1)/2}},\end{aligned} \tag{21}$$

- where $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ is the beta function and $Y^{**}_{i_0,m}$, as defined in (18), is the moving sum squares for the transformed data $\{X_i^{**} = X_i/R; 1 \le i \le M\}$, defined in (16).
- The final representation of the conditional GLRT statistic depends only on the joint distribution of $\{X_i^{**} = X_i/R; 1 \le i \le M\}$, which under H_0, does not depend on the unknown value of σ_0, and under H_1, depends only on σ_1/σ_0.
- Moreover, for a fixed window size m, the function $g(Y^{**}) = (Y^{**})^{(m-1)/2}(1-Y^{**})^{(M-m-1)/2}$ is decreasing in Y^{**}

Multiple and Variable Window Scan Statistics

- Therefore, the conditional GLRT, Λ is increasing in Y^{**}
- Hence, for a known fixed value of m, the conditional GLRT will be the same as our fixed window scan statistic, conditional on the sufficient statistic for σ_0^2, discussed earlier.
- Zhao and Glaz (2016b) present an algorithm that finds LR^*, the value of the conditional $GLRT$ statistic; m^* is the most likely window size where a possible upward local change in variance has occurred and $i_0^*(m^*)$ is the most likely starting location for the local change. We refer to the conditional GLRT statistic, Λ, as a *conditional variable window scan statistic*.
- The p-value for Λ can be obtained by performing N simulation runs, each having M iid $N(0,1)$ random variables, and then repeating the algorithm in Zhao and Glaz (2016b) for each of the simulated M-sequences.
- Numerical results to compare power of the conditional multiple window scan statistic P_{min}, conditional variable variable window scan statistic Λ and conditional fixed window scan statistic $S^{**}_{m,M}$.

Scan Statistics for Two Dimensional Data

- The results that have been discussed above have been extended to two dimensional data
- Zhao and Glaz (2017). Scan Statistics for Detecting a Local Change in Variance for Two Dimensional Normal Data. *Commun. Stat. Theor. Meth. Ser.* A, Vol. 46, No. 11, 5517-5530.
- The references for the one dimensional case:
- Zhao, B. and Glaz, J. (2016a). Scan Statistics for Detecting a Local Change in Variance for Normal Data.with Known Variance. *Methodology and Computing in Applied Probability* **18**, 967-978.
- Zhao, B. and Glaz, J. (2016b). Scan Statistics for Detecting a Local Change in Variance for Normal Data.with Unknown Variance. *Statistics and Probability Letters* **110**, 137-145.