CHANGEPOINT DETECTION USING NONPARAMETRIC PROCEDURES

Sivanandan Balakumar

Abstract. This study investigates the detection of a changepoint due to location or scale or both location and scale changes in a sequence of univariate random variables. Two changepoint models, namely, the single change and the continuous change are considered and the appropriate test statistics are derived. The proposed test statistics are functions of two linear rank statistics, of which one is odd-translation invariant, sensitive to location shift and the other one is even-translation invariant, sensitive to scale shift. The asymptotic distributions of the proposed statistics are obtained under the null hypothesis.

1. Introduction. In this paper a nonparametric procedure for the problem of testing for location or scale or both location and scale changes occurring at an unknown time point is considered. In a sequence of independent random variables $X_1, \ldots, X_N$ with continuous and differentiable distribution functions (d.fs.), $F(x, \theta_i)$ for $i = 1, \ldots, N$; where the θ_i's are either location or scale parameters, if the random variables X_i, $i = 1, \ldots, n$ for $1 \leq n < N$ have d.fs. $F(x, 0)$ and the random variables X_i, $i = n+1, \ldots, N$ have d.fs. $F(x, \theta)$ with $\theta \neq 0$ then the sequence X_i, $i = 1, \ldots, N$ is said to have a single changepoint at time point n. The change in distribution may occur continuously over a period of time. This type of change may be referred to as a continuous change.

Changepoint problems have been studied quite extensively by various authors, using parametric and nonparametric procedures. In some of the studies the authors have assumed known underlying distributions. However, a good amount of work has also been done for the case of arbitrary underlying distributions. In these studies, changes in location or scale have been considered. To my knowledge, nothing has been done with respect to testing for location and scale changes together in a changepoint setting. Lombard [3] proposed rank statistics for detecting single abrupt, multiple abrupt, and smooth changepoints.

In this study, statistics that are functions of two linear rank statistics, one sensitive to location shift and the other one sensitive to scale shift are proposed. In section 2, the single changepoint model (SCM) is defined and the corresponding test statistic is obtained. The asymptotic distribution of the test statistic under null hypothesis is also considered. The same information for the continuous changepoint model (CCM) are given in section 3.

2. Single Changepoint Model and Test Statistic. Let $X_1, \ldots, X_N$ be a sequence of independent random variables with continuous d.fs. $F[(1 + a_i)x + b_i]$,

$i = 1, \ldots, N$ where a_i's and b_i's are the scale and location parameters, respectively. The changepoint model SCM may be defined as follows.

$$(1) \qquad a_i = \begin{cases} 0, & 1 \le i \le n \\ \text{a}, & n+1 \le i \le N, \end{cases} \qquad b_i = \begin{cases} 0, & 1 \le i \le n \\ \text{b}, & n+1 \le i \le N. \end{cases}$$

The accompanying null and alternative hypotheses are:

$$H_0 \ : \ a = 0 \text{ or } b = 0 \text{ or both}$$

and

$$H_a \ : \ a \neq 0 \text{ or } b \neq 0 \text{ or both.}$$

In this case, if the changepoint is known then the problem becomes the standard two-sample problem and the test statistic is readily obtained and is given by

$$S_n = \sum_{i=n+1}^{N} a_N(R_i),$$

where $a_N(R_i)$'s are the score functions based on the ranks R_i's of X_i's. In the model above, since the changepoint n is unknown, the required statistic takes the form

$$S_N = \sum_{n=1}^{N-1} \sum_{i=n+1}^{N} a_N(R_i) = \sum_{i=1}^{N} (i-1) a_N(R_i).$$

Now we may use appropriate score functions to yield odd-translation invariant and even-translation invariant statistics that are suitable for testing location and scale changes, respectively. Thus, the test statistics can be formulated as

$$(2) \qquad S_{Nk} = \sum_{i=1}^{N} c_i a_{Nk}(R_i), \quad k = 1, 2,$$

where S_{N1} is odd-translation invariant, S_{N2} is even-translation invariant and $c_i = i - 1$, for $i = 1, \ldots, N$. The statistics S_{Nk}, $k = 1, 2$ are suitable for detecting

location and scale changes respectively. To formulate a statistic for testing location and scale changes, consider the following definitions. Let

$$S^T = (S_{N1}, S_{N2}), \quad \mu^T = (\mu_{N1}, \mu_{N2}), \quad A = \text{diag}\ (\sigma^2_{N1}, \sigma^2_{N2}) \tag{3}$$

where

$$\mu_{Nk} = E[S_{Nk}] = N(N-1)/2 \int_0^1 \phi_k(u)du$$

$$\sigma^2_{Nk} = \text{Var}[S_{Nk}] = N(N^2-1)/12 \int_0^1 [\phi_k(u) - \overline{\phi}_k]^2 du, \quad \text{and}$$

$$\zeta^2_{Nk} = \text{Cov}\ [S_{N1}, S_{N2}] = N(N^2-1)/12 \int_0^1 [\phi_1 - \overline{\phi}_1][\phi_2 - \overline{\phi}_2]du.$$

The score functions $a_{Nk}(\cdot)$, $k = 1, 2$ are given by square integrable functions ϕ_k, $k = 1, 2$ defined as

$$a_{Nk}(R_i) = \phi_k(R_i/N + 1). \tag{4}$$

Now, for testing location and scale simultaneously in a changepoint problem the required statistic is given by

$$L_1 = (S - \mu)^T A^{-1} (S - \mu). \tag{5}$$

The null hypothesis in the model SCM given by (1) is rejected for a large value of the statistic L_1. The following theorem gives the asymptotic distribution of L_1 under H_0.

Theorem 1. When H_0 holds, then

$$\lim_{N\to\infty} P[L_1 \le x] = P[X_2^2 \le x], \ x > 0,$$

where X_2^2 is a Chi-square random variable with two degrees of freedom.

Proof. The statistic L_1 can be written as

$$L_1 = \sum_{k=1}^{2} [(S_{Nk} - \mu_{Nk})/\sigma_{Nk}]^2.$$

Since for each k, S_{Nk} is a linear rank statistic whose regression constants c_i's satisfy the well known Noether's condition, it follows from Theorem V.1.5a of Hajek and Sidak [2] that $(S_{Nk} - \mu_{Nk})/\sigma_{Nk}$ for each k has an asymptotic standard normal distribution. Furthermore, since the statistics S_{N1} and S_{N2} are uncorrelated, they are asymptotically independent. Therefore the statistic L_1 converges in distribution to $Z_1^2 + Z_2^2$ where Z_1 and Z_2 are independent standard normal random variables. Hence, the result of the theorem.

Similar results can be obtained for the continuous changepoint problem and is given in the following section.

3. Continuous Changepoint Model and Test Statistic. The continuous changepoint model (CCM) is defined as follows

$$a_i = \begin{cases} 0, & 1 \le i \le n \\ \text{a(i-n)/(N-n)} & n+1 \le i \le N, \end{cases} \qquad b_i = \begin{cases} 0, & 1 \le i \le n \\ \text{b(i-n)/(N-n)} & n+1 \le i \le N \end{cases}$$

with the hypotheses

$$H_0 \ : \ a = 0 \text{ or } b = 0 \text{ or both}$$

and

$$H_a \ : \ a \neq 0 \text{ or } b \neq 0 \text{ or both.}$$

The method of Hajek and Sidak [2] can be used to derive a test statistic for this changepoint model with known n, as

$$T_n = -\sum_{i=n+1}^{N} (N-i)\ a_N(R_i) + (N-n) \sum_{i=n+1}^{N} a_N(R_i) = \sum_{i=n+1}^{N} (i-n)\ a_N(R_i).$$

Now for the CCM with unknown n, the required statistic is obtained as

$$T_N = \sum_{n=1}^{N-1} \sum_{i=n+1}^{N} (i-n)\ a_N(R_i) = \sum_{i=1}^{N} d_i\ a_N(R_i),$$

where $d_i = i(i-1)/2$. The statistic T_N may be modified as in the case of SCM, to test for location and scale changes as follows

$$T_{Nk} = \sum_{i=1}^{N} d_i\ a_{Nk}(R_i),\ k = 1, 2 \tag{6}$$

so that T_{N1} is odd-translation invariant and T_{N2} is even-translation invariant. The statistics T_{N1} and T_{N2} are suitable for detecting location and scale changes, respectively.

In order to detect both the location and scale changes, like in the SCM the following statistic can be considered.

$$L_2 = (T - v)^T B^{-1} (T - v), \tag{7}$$

where

$$T^T = (T_{N1}, T_{N2}), \quad v^T = (v_{N1}, v_{N2}),$$

$$v_{Nk} = E[T_{Nk}] = N(N^2 - 1)/6 \int_0^1 \phi_k(u) du, \quad B = \text{diag } (\tau_{N1}^2, \tau_{N2}^2), \text{ and}$$

$$\tau_{Nk}^2 = \text{var } [T_{Nk}] = N(N^2 - 1)(4N^2 - 1)/180 \int_0^1 [\phi_k(u) - \overline{\phi}_k]^2 du \quad \text{for } k = 1, 2$$

and $\phi_k(\cdot)$ is as defined in (4). And it can be shown that

$$\text{Cov } [T_{N1}, T_{N2}] = N(N^2 - 1)(4N^2 - 1)/180 \int_0^1 [\phi_1 - \overline{\phi}_1][\phi_2 - \overline{\phi}_2] du.$$

With these definitions and results we use the procedures of section 2 to come up with the following results regarding the asymptotic distributions of the test statistic L_2.

Theorem 2. When H_0 holds, then

$$\lim_{N \to \infty} P[L_2 \leq x] = P[X_2^2 \leq x], \ x > 0,$$

where X_2^2 is the random variable with a Chi-square distribution with two degrees of freedom.

4. Concluding Remarks. The detection of single and continuous change-points in the presence of location and scale changes is studied in this paper. A combination of odd-translation invariant and even-translation invariant linear rank statistics is used in this case. Similar procedures to the one used can be applied

in detecting multiple changepoints in a sequence of random variables. It is also believed that a similar procedure may be applied in detecting a changepoint in a sequence of exchangeable random variables. Note that exchangeability does not imply independence.

An extension of the procedure mentioned in this paper can be carried out effectively for multivariate changepoint problems. The statistics L_1 and L_2 are of the form used by Duran, Tsai and Lewis [1] for simultaneous location-scale testing on two-sample problems. Also note that Lombard [3] used similar models.

Acknowledgement. I am grateful to the referees for their helpful suggestions.

References

1. B. S. Duran, W. S. Tsai, and T. O. Lewis, "A Class of Location-Scale Nonparametric Tests," *Biometrika*, 63 (1976), 173–176.

2. T. Hajek and Z. Sidak, *Theory of Rank Tests*, Academic Press, New York, 1967.

3. F. Lombard, "Rank Tests for Changepoint Problems," *Biometrika*, 74 (1987), 615–624.

Sivanandan Balakumar
Department of Natural Sciences and Mathematics
Lincoln University
Jefferson City, MO 65102
email: balakuma@lincolnu.edu

æ