AD-A198 406

Unclassified

DTIC FILE COPY

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 88-0796

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): Detection of Change Points Using Rank Methods

5. TYPE OF REPORT & PERIOD COVERED: Report ~~Technical~~ - March 1988

6. PERFORMING ORG. REPORT NUMBER: 88-02

7. AUTHOR(s): B. Q. Miao and L. C. Zhao

8. CONTRACT OR GRANT NUMBER(s): AFOSR ~~AFSC~~-88-0030

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Center for Multivariate Analysis
515 Thackeray Hall
University of Pittsburgh, PA 15260

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304 A6

11. CONTROLLING OFFICE NAME AND ADDRESS:
Air Force Office of Scientific Research/ BLDG 410
Dept. of the Air Force
Bolling Air Force Base, DC 20332

12. REPORT DATE: March 1988

13. NUMBER OF PAGES: 13

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same as 11.

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DTIC ELECTE AUG 26 1988 D

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Change point, Detection, Directional data, Nonparametric method, rank statistics.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this paper, the detection and estimation of change points of local parameters are studied by means of localization procedures and rank statistics. These techniques are also applied to detection and estimation of the change points of scale parameters and that of location parameters of directional data.

DD FORM 1473, 1 JAN 73

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

88 8 25 146

AFOSR-TR- 88-0796

Center for Multivariate Analysis

University of Pittsburgh

88 8 25 140

DETECTION OF CHANGE POINTS USING RANK METHODS

B. Q. Miao and L. C. Zhao

Center for Multivariate Analysis
University of Pittsburgh

Technical Report No. 88-02

February 1988

Center for Multivariate Analysis
Fifth Floor Thackeray Hall
University of Pittsburgh
Pittsburgh, PA 15260

Research Sponsored by the Air Force Office of Scientific Research under Grant AFOSR - 88 - 0030. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

i

DETECTION OF CHANGE POINTS USING RANK METHODS

B.Q. Miao and L. C. Zhao

Center for Multivariate Analysis
University of Pittsburgh
Pittsburgh PA 15260

AMS 1980 Subject Classifications: Primary 62G30, Secondary 62D05.

Key words and phrases: Change point, Detection, Directional data, Nonparametric method, rank statistics.

ABSTRACT

In this paper, the detection and estimation of change points of local parameters are studied by means of localization procedures and rank statistics. These techniques are also applied to detection and estimation of the change points of scale parameters and that of location parameters of directional data.

1. INTRODUCTION

Change point problem arises in many fields and attracts the attention of many authors. The techniques employed to detect and estimate the change point can generally be classified into two categories : parametric (Krishnaiah and Miao .1988), and nonparametric, (Csörgö and Horváth, 1988). Bayesian methods also plays a major role, (Zacks, 1983).

In this paper, ~~we~~ concentrate ~~our~~ attention on the problem of detection of change points of location parameter by localization and rank statistics when data are large. Our method is different from, and has some advantages over, the existing methods, such as CUSUM (cumulative sum) and Csörgö and Horváth's non-sequential nonparametric AMOC (at most one change) procedures. First, localized procedures reduce computation. Second, these detecting and estimating procedures require no moment condition, instead ~~we~~ only assume that observed data come from a continuous distribution with a unique median.

2. MODEL AND DETECTING AND ESTIMATING PROCEDURE

Let $x(t)$ be an independent process on the interval $(0,1]$ whose marginal distributions differ only by location parameters. Specifically speaking, there exist $t_0, t_1, \ldots, t_{q+1}$ and $a_1, \ldots, a_{q+1}$ such that $0 = t_0 < t_1 < \ldots < t_q < t_{q+1} = 1$ and

$$x(t) \sim F(x-a_j), \quad \text{if} \quad t_{j-1} < t \leqq t_j,\ j = 1,2,\ldots,q+1, \tag{2.1}$$

with

$$a_j \neq a_{j+1} \quad \text{for} \quad j = 1,\ldots q. \tag{2.2}$$

As usual, $t_1,\ldots,t_q$ are called change points. In practical applications, the number q and the locations of change points are unknown. To estimate q and $t_1,\ldots t_q$, we sample this process sequentially at equal distances, and get $x(1/n), x(2/n),\ldots x(n/n)$. By assumption, $x(1/n),\ldots,x(n/n)$ are independent. Define $\overset{\circ}{k}_0 = 0$, $\overset{\circ}{k}_{q+1} = n$, and $\overset{\circ}{k}_1, \ldots, \overset{\circ}{k}_q$ as follows:

$$|\overset{\circ}{k}_i(n)/n - t_i| < 1/n, \quad i = 1,\ldots,q. \tag{2.3}$$

Then we have

$$x(i/n) \sim F(x-a_j), \quad \text{if} \quad \overset{\circ}{k}_{j-1}(n) < i \leqq \overset{\circ}{k}_j(n), \qquad j = 1,2,\ldots,q+1, \tag{2.4}$$

From this fact, to estimate $(t_1,\ldots,t_q)$ is equivalent to estimating $(\overset{\circ}{k}_1(n)/n,\ldots,\overset{\circ}{k}_q(n)/n)$. For simplicity, set $x(i/n) = x_i$ and $\overset{\circ}{k}_i(n) = \overset{\circ}{k}_i$. Hereafter, $\alpha_n \gg \beta_n$ means $\alpha_n/\beta_n \to \infty$. Let $m = m_n$ and C_n be positive integers such that $n \gg m_n \gg C_n \gg \log n$.

Set $A_{k,m} = \{x_{k-m+1}, \ldots, x_k, x_{k+1}, \ldots, x_{k+m}\}$, $k = m,\ldots,n-m$, k_j be the rank of x_j in $A_{k,m}$. Define

$$y_{k,j} = \begin{cases} 1 \text{ if } & x_{k-m+j} \leqq m, \\ -1 & \text{otherwise} \end{cases} \quad j = 1,\ldots,m. \tag{2.5}$$

$$S_{k,m} = \sum_{j=1}^{m} y_{k,j}\ . \tag{2.6}$$

$$D_n = \{k: k = m,\ldots,n-m : S^2_{k,m}/m > C_n\}, \tag{2.7}$$

$$k_{1,n} = \min\{k : k \in D_n\},$$

$$D_{1,n} = \{k : k \in D_n,\ \ k - k_{1,n} < 3m\}.$$

Next, put

$$k_{2,n} = \min\{k : k \in D_n - D_{1,n}\}$$

$$D_{2,n} = \{k : k \in D_n - D_{1,n}, k\text{-}k_{2,n} < 3m\}.$$

Continuing this process, we can define $D_{2,n}$,$D_{2,n}$ ..., which are easily seen to be nonempty. We have

$$D_n = D_{1,n} + D_{2,n} + \ldots + D_{\hat{q},n}.$$

Define

$$\hat{t}_j = 2^{-1}[k_{j,n} + \max(k : k \in D_{j,n})],\ \ j = 1,2,\ldots,\hat{q}.$$

We have the following theorem.

<u>THEOREM</u>. If the distribution F is continuous and has unique median, then $(\hat{q}, \hat{t}_1, \ldots, \hat{t}_{\hat{q}})$ is a strongly consistent estimate of $(q, t_1, \ldots, t_q)$.

3. LEMMAS

To prove the above theorem , we need some lemmas.

LEMMA 1. (Hoeffding, 1963) Let the population C consists of N values $C_1,\ldots,C_N$. Let $x_1,\ldots,x_n$ denote a random sample without replacement from C and let $y_1,\ldots,y_n$ denote a random sample with replacement from C. If the function f(x) is continuous and convex then

$$Ef(\sum_{i=1}^{n} x_i) \leqq Ef(\sum_{i=1}^{n} y_i).$$

LEMMA 2. Let the notation be defined as in section 2 and define

$$\mathcal{B}_{k,m} = \{x_{k-m+1},\ldots,x_k\}, \quad k = m, m+1,\ldots,n. \tag{3.1}$$

$$S_{k,m} = \sum_{j=1}^{m} y_{k-m+j}, \tag{3.2}$$

$$\mathcal{B}_n = \{k : \exists\, j,\ 1 \leqq j \leqq q+1, \text{ such that } k^o_{j-1} + 1 \leqq k - m + 1 < k \leqq k^o_j\}. \tag{3.3}$$

Then, we have

$$S^2_{k,m}/m = 0(\log n) \quad \text{a.s.}$$

uniformly for all $k \in \mathcal{B}_n$.

Proof. Let $z_{k,1},\ldots,z_{k,m}$ be a random sample with replacement from population $\{1,\ldots,1,-1,\ldots,-1\}$, where the number of 1's and -1's are both m. By lemma 1 for any $t \in (0,1/4)$ and $A > 0$, we have

$$P\{S_{k,m} \geqq A\sqrt{m\log n}\} \leqq \exp\{-tA\sqrt{m\log n}\} Ee^{t\sigma_{k,m}}$$

$$\leqq \exp\{-tA\sqrt{m\log n}\} E \exp\{t\sum_{j=1}^{m} z_{k,j}\}$$

$$= \exp\{-tA\sqrt{m\log n}\}\,(E\exp\{tz_{k,j}\})^m$$

$$= \exp\{-tA\sqrt{m\log n}\}((e^t + e^{-t})/2)^m$$

$$\leqq \exp\{-tA\sqrt{m\log n} + mt^2 e^t/2\}.$$

Since $m >> \log n$, it is possible for n large to take $t = A\sqrt{(\log n)/m} < 1/4$. It follows

$$P(S_{k,m} \geqq A\sqrt{m\log n}) < \exp\{-0.3A^2\log n\}. \tag{3.4}$$

A similar argument gives

$$P(\sigma_{k,m} \leqq -A\sqrt{m\log n}) < \exp\{-0.3A^2\log n\}. \tag{3.5}$$

The inequalities (3.3) and (3.4) imply

$$\sum_{n=1}^{\infty} P(\sup_{k\in B_n} |S_{k,m}| \geqq A\sqrt{m\log n}) \leqq \sum_{n=1}^{\infty} 2n\exp\{0.3A^2\log n\}.$$

This is a convergence series if $0.3A^2 > 2$, Take $A = 3$, by Borel-Cantelli lemma, with probability one for large n, we have uniformly for all $k \in \mathcal{B}_n$,

$$|S_{k,m}| \leqq 3\sqrt{m\log n},$$

$$S^2_{k,m}/m \leqq 9\log n \qquad \text{a.s.}$$

uniformly for all $k \in \mathcal{B}_n$.

LEMMA 3. Let $x^*_{i,k}$ denote the i-th sample order statistic in $\mathcal{B}_{k,m}$. Let $[\alpha m]$ denote the integer part of αm. Assume that the α-quantile of the continuous distribution F is unique, and is denoted by μ_α. Then

$$x^*_{[\alpha m],k} \to \mu_\alpha + a_j, \quad \text{a.s.}$$

uniformly for all $k \in \mathcal{B}_n$, where $k^o_{j-1} + 1 \leqq k - m + 1 < k \leqq k^o_j$.

Proof. Without loss of generality, we assume that $0 < \alpha < 1$, $k^o_{j-1} + 1 \leqq k - m + 1 < k \leqq k^o_j$ for some fixed j and $a_j = 0$. Write $r = [\alpha m]$, then for any $\varepsilon > 0$,

$$P(|x^*_{[\alpha m],k} - \mu_\alpha| \geqq \varepsilon) = P(|x^*_{r,k} \leqq \mu_\alpha - \varepsilon) + P(x^*_{r,k} \geqq \mu_\alpha + \varepsilon)$$

$$\triangleq I_1 + I_2.$$

Set $F(\mu_\alpha - \epsilon) = \alpha - \delta$. By the uniqueness of the α - quantile of F, $\delta > 0$. Without loss of generality, we can assume that $\alpha - \delta > 0$. Since F(x) is continuous, one gets,

$$I_1 = P(x^*_{r,k} \leqq \mu_\alpha - \varepsilon)$$

$$= \frac{m!}{(r-1)!(m-r)!}\int_0^{F(\mu_\alpha - \varepsilon)} t^{r-1}(1-t)^{m-r}dt$$

$$= \frac{r.m!}{r!(m-r)!}\int_0^{\alpha-\delta} t^{r-1}(1-t)^{m-r}dt$$

For a fixed $\theta \in (0,1)$, put

$$g_\theta(t) = t^\theta(1-t)^{1-\theta}.$$

It is easy to see that g(t) is increasing in t on the interval $[0, \theta]$, which implies that $t^{r-1}(1-t)^{m-r}$ is increasing for $t \in [0, (r-1)/(m-1)]$. Since $r = [\alpha m]$, we get $\alpha - \delta < (r-1)/(m-1)$ for large m. Using Stirling's formula, we obtain

$$I_1 \leqq \frac{rm!}{r!(m-r)!}(\alpha - \delta)^{r-1}(1-\alpha+\delta)^{m-r}$$

$$\leqq \frac{C_1 m\sqrt{2\pi m}}{\sqrt{2\pi . 2\pi . r(m-r)}} \frac{(\alpha-\delta)^{r-1}(1-\alpha+\delta)^{m-r}}{(r/m)^r((m-r)/m)^{m-r}}$$

$$\leqq C\sqrt{m}\, q_1^m \quad , \tag{3.6}$$

where C_1 and C are constants depending on α, δ, and

$$q_1 = \{(\alpha-\delta)/(\alpha-\delta/2)\}^{\alpha}\{(1-\alpha+\delta)/(1-\alpha+\delta/2)\}^{1-\alpha},$$

$$= g_\alpha(\alpha-\delta)/g_\alpha(\alpha-\delta/2) < 1.$$

Similarly, one can show that

$$P(x^*_{r,k} \geqq \mu + \varepsilon) \leqq C\sqrt{m}\, q_2^m, \tag{3.7}$$

where $0 < q_2 < 1$. Take $q = \max(q_1,q_2)$, then (3.6), (3.7) and $m_n >> \log n$ togather imply

$$\sum_{n=1}^{\infty} P(\sup_{k \varepsilon B_n} | x^*_{[\alpha m],k} - \mu_\alpha - a_j | \geqq \varepsilon) \leqq \sum_{n=1}^{\infty} 2cn\sqrt{m}\, q^m < \infty.$$

By Borel-Cantelli lemma, it follows that

$$x^*_{[\alpha m],k} \to \mu_\alpha + a_j, \quad \text{a.s.}$$

uniformly for all $k \in \mathcal{B}_n$, where $k^o_{j-1} + 1 \leqq k - m + 1 < k \leqq k^o_j$.

<u>LEMMA</u> 4. Let $x_i \in A_{k^o_j,m}$, $j = 1,2,\ldots,q$, $S_{k^o_j,m}$ be as defined by (2.6). Then there exists a positive λ such that with probability one for n large,

$$S^2_{k^o_j,m}/m \geqq m\lambda^2, \quad j = 1,\ldots,q.$$

Proof. Let $x_i \in A_{k^o_j,m}$ for some fixed j. Without loss of generality, we assume $a_{j+1} > a_j$. Take $\lambda \,\varepsilon\, (0,1/2)$ satisfying the following conditions:

1° the $(1/2-\lambda)$ and $(1/2+\lambda)$ - quantiles of the continuous distribution $F(x)$ are unique.

2° $\mu_{1/2} - \mu_{1/2-\lambda} < (a_{j+1} - a_j)/2,$ (3.8)

$$\mu_{1/2+\lambda} - \mu_{1/2} < (a_{j+1}-a_j)/2. \tag{3.9}$$

Note that λ exists since the median of F is unique and the set $\{p:0 < p < 1$, the p-quantile is not unique $\}$ is countable.

Let x^*_{i,k_j° and y^*_{i,k_j° denote the i-th order statistics in $B_{k_j^\circ, m}$ and $B_{k_j^\circ+m,m}$ respectively. By lemma 3, for all k_j°, $1 \leqq j \leqq q$,

$$x^*_{[m/2+m],k_j^\circ} \to \mu_{(1/2)+\lambda} + a_j, \quad \text{a.s.} \tag{3.10}$$

$$y^*_{[m/2-m],k_j^\circ} \to \mu_{(1/2)-\lambda} + a_{j+1}, \quad \text{a.s.} \tag{3.11}$$

By (3.8) and (3.9), with probability one for n large,

$$x^*_{[m/2+m],k_j^\circ} < y^*_{[m/2-m],k_j^\circ}, \quad \text{a.s.} \tag{3.12}$$

Now consider the order statistics in the combined sample $B_{k_j^\circ,m} \cup B_{k_j^\circ+m,m} = A_{k_j^\circ,m}$. By (3.12), and the fact that $[m/2 + \lambda m] + [m/2 - \lambda m] \leqq m$, at least $[m/2 + \lambda m]$ x_i's can be found in $B_{k_j^\circ,m}$ which are the first m order statistics of $A_{k_j^\circ,m}$. Therefore, with probability one for n large, we have

$$S_{k_j^\circ,m} \geqq [m/2 + \lambda m] - [m/2 - \lambda m] \geqq \lambda m \quad . \tag{3.13}$$

The lemma is proved.

4. PROOF OF THEOREM AND SOME APPLICATIONS.

Proof of Theorem.

Let k_j^o be defined by (2.3). For fixed j, by lemma 4, we have, with probability one for n large,

$$S^2_{k_j^o,m}/m \geqq m\lambda^2.$$

By the definition of D_n, $k_j^o \in D_n$, $j = 1,...,q$. Further, by the definition of $D_{j,n}$'s and $n >> m$, it follows that with probability one for large n, $k_1^o,...,k_q^o$ belong to different $D_{j,n}$'s, which in turn implies

$$\hat{q} \geqq q \quad . \tag{4.1}$$

with probability one for large n. On the other hand, from Lemma 2 and $C_n >> \log n$, it follows that with probability one for n large $S^2_{k,m}/m < C_n$. Further take $0 < \varepsilon < 1/2$ and write

$$K_n = \{k : m \leqq k \leqq n - m, |k/n - t_j| \geqq (m/n)(1+\varepsilon) \quad \text{for } j = 1,...,q\}. \tag{4.2}$$

Then, since $n >> m_n$, and $A_{k,m}$ consists of iid. random variables, it follows that with probability one for n large

$$D_n \cap K_n = \Phi. \tag{4.3}$$

By the definition of $D_{j,n}$ and the fact that $2m(1+\varepsilon) < 3m$, we have with probability one for large n

$$\hat{q} \leqq q. \tag{4.4}$$

Combining (4.1) to (4.4), it follows that $\hat{q} = q$ with probability one for n large, and

$$\lim_{n\to\infty} \hat{t}_j = t_j \qquad j = 1,2,...,q.$$

Thus, the theorem is established.

Our procedure can also be applied to the detection of change points of scale parameters. What we need to do is merely to consider $\log |x(t)|$ instead of $x(t)$.

Likewisely this method can be applied to the detection of change points of location parameters in a circle, i.e. directional data. Lombard (1986) proposed this problem and discussed the related testing and estimating problem when the number of change points is known. Our technique can handle the case in which this number is unknown.

For directional data, fix an origin and an axis and let $\{x(t), 0 \leqq t \leqq 1\}$ be the angle process, where $0 \leqq x(t) < 2\pi$. Assume that $0 = t_0 < t_1 < \ldots < t_q < t_{q+1} = 1$, $t_1,\ldots,t_q$ are the change points, so that $x(t) \sim F(x-a_j)$ if $t_{j-1} < t < t_j$, $j = 1,\ldots, q+1$, where $a_j \not\equiv a_{j+1}$ (mod 2π), $j = 1,\ldots,q$, $a_1,\ldots a_{q+1}$ denote the angular location parameters. The speciality of the angle process is that the angle is always in $[0,2\pi)$. So the angle of $\alpha + \beta$ equals $\alpha + \beta - 2\pi$ if $0 \leqq \alpha,\beta < 2\pi$ and $\alpha + \beta \geqq 2\pi$. A closer look at our proof reveals that the speciality has an effect only on (3.10) and (3.11) when $\mu_{(1/2+\lambda)} + a_j = 0$ (mod 2π) or $\mu_{(1/2-\lambda)} + a_{j+1} = 0$ (mod 2π) for some j, $1 \leqq j \leqq q$. Therefore, if continuous distribution $F(x)$ has a unique median and $\mu_{1/2} + a_j \not\equiv 0$ (mod 2π) for all j, $1 \leqq j \leqq q$, our procedure can be used to detect and estimate the change points of directional data. In practical applications, the origin and axis are chosen after gathering the data, in a way to make the analysis more convenient.

BIBLIOGRAPHY

[1] Csörgö, M. and Horváth, L. (1988), Nonparametric methods for change point problems. *Multivariate Analysis* VII. to appear.

[2] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association,* Vol.58, 13-30.

[3] Krishnaiah, P. R. and Miao, B. Q. (1988). Review about estimates of change point. *Multivariate Analysis* VII. to appear.

[4] Lombard, F. (1986). The change-point problem for angular data: a nonparametric approach. *Technometrics*, Vol. 28, No. 4, 391-397.

[5] Pettitt, A. N. (1980). A simple cumulative sum type statistics for the change-point problem with zero-one observation. *Biometrika* Vol.67, 79-84.

[6] Zacks, S. (1983). Survey of classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures of testing and estimation. Recent advances in statistics, ed. M. H. Rizvi, New York: Academic Press, 245-269.