A Likelihood Integrated Method for Change-point Problem with Directional Data

Ashis SenGupta
Indian Statistical Institute, Kolkata. INDIA

and

Arnab Kumar Laha
Indian Institute of Management, Ahmedabad, INDIA

ABSTRACT

In this paper we introduce a new likelihood-based method, which we call the likelihood integrated method. This method is distinct from the well known integrated likelihood method. We consider the non-trivial problem of change-point with directional data. We demonstrate that here this new method yields a novel and simple exploratory graphical analysis, which can be easily implemented in practice. The method is applied to analysis of two real life data sets. The results found by this simple method are seen to be quite similar to those obtained earlier by different complicated formal methods.

Key Words & Phrases: Directional data, Change-point problem, Circular normal distribution, Papakonstantinou's distribution, Likelihood integrated method.

Introduction

The problem of detection of a change in the distribution in a given finite sequence of independent observations is commonly referred to as the change -point problem. This problem was introduced in the statistical literature by Page(1955) in the context of detection of abrupt change in process parameters which leads to poor quality products. The now well-known Cusum chart is one of the earliest techniques suggested to deal with this problem. The change-point problem has received considerable attention in the linear case. The two main streams of work pertain to the parametric set-up with normal distribution as the underlying distribution and the non-parametric set-up see eg, Chernoff and Zacks (1964), Hinkley (1970), Sen and Srivastava (1973, 1975a, 1975b), Chen and Gupta (2000) etc. In the context of circular or directional data, the change-point problem arises in many applications, e.g. detection of time of change of wind direction, direction of movements of icebergs, propagation of cracks etc. In the circular case not much work has been done on this problem. Lombard (1986) initiated work in the context of directional data in the non-parametric frame work while Laha (2001) discusses the change-point problem in the parametric framework with the circular normal distribution as the underlying distribution of the data.

In this paper we discuss the change-point problem for angular variables by using a new approach which we call the “likelihood integrated” approach to distinguish it from the well known integrated likelihood approach (Berger et. al., 1999). In the usual integrated likelihood approach the likelihood based on all the observations is integrated with respect to the conditional (joint) prior distribution of the nuisance parameter(s) given the parameter(s) of interest. The (joint) prior distribution is assumed to be absolutely continuous with respect to the Lebesgue measure and hence is assumed to have a density. The integration is carried out over all possible values of the nuisance parameter(s). This process eliminates the nuisance parameter(s) and the resultant quantity is called the integrated likelihood which can be used for inference on the parameters of interest. Berger et. al. (1999) give an excellent discussion of the integrated likelihood approach.

In our approach, we assume that the observations are mutually independent. We first obtain the integrated likelihood for one observation by integrating the likelihood for one observation with respect to the conditional (joint) prior distribution of the nuisance parameter(s) given the parameter(s) of interest over all possible values of the nuisance parameter(s). Thus the resulting expression does not involve the nuisance parameter(s). The "likelihood integrated" is then formed by taking the product of the integrated likelihood for each observation. It is evident that the likelihood integrated method is less general than the integrated likelihood method since it assumes that the observations are independent. However, this is not a very serious restriction as the independence of observations is a widely prevalent assumption in statistics. We illustrate the method for two circular distributions – the circular normal (also called von Mises) distribution and the Papakonstantinou's distribution.

The circular normal distribution is the most popular distribution for directional data. It is a symmetric unimodal distribution having probability density function (p.d.f.)

$$f(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)}\ ,\ 0\le\mu<2\pi\ ,\ \kappa>0$$

where $I_0(\kappa)$ is the modified Bessel function of order 0. If the angular random variable Θ follows the circular normal distribution with parameters μ and κ then we write $\Theta\sim CN(\mu,\kappa)$. The parameter μ is called the mean direction (also modal direction) and the parameter κ is called the concentration parameter. For more details on the circular normal distribution the reader may look into Jammalamadaka and SenGupta (2001).

The Papakonstantinou's distribution is a skew circular distribution having the p.d.f.

$$f(\theta;\nu,k)=\frac{1}{2\pi}+\frac{k}{2\pi}\sin(\theta+\nu\sin\theta)\,,0\le\theta<2\pi,\,-1<k<1\,.$$

If the angular random variable Θ follows the Papakonstantinou's distribution with parameters ν and k then we write $\Theta\sim P(\nu,k)$. It may be noted that for $k=0$ this distribution reduces to the circular uniform distribution (which incidentally, is a symmetric distribution) and for $k\neq 0$ we obtain a skew circular distribution. Further details about this distribution can be found in Batschelet (1981).

We apply the likelihood integrated method for analysis of two real life data sets – flare data (Lombard, 1986) and wind data (Weijers et. al., 1995). We discuss a simple graphical method for exploratory analysis of data sets having at most one change point. A method for identifying the change point is also discussed. The results obtained from these analyses demonstrate the potential of this approach for exploratory analysis of directional data suspected to have a change point.

Likelihood Integrated Method for Change-point Problem

Let $\Theta_1, \Theta_2, ..., \Theta_n$ be independent angular observations. Suppose the p.d.f. of Θ_i is $f(\theta; \boldsymbol{\varphi}_i, \boldsymbol{\eta})$ where $\boldsymbol{\varphi}_i$ is the vector of the parameters of interest and $\boldsymbol{\eta}$ is the vector of nuisance parameters. Let $\pi(\boldsymbol{\eta} \mid \boldsymbol{\varphi}_i)$ be the conditional joint prior density of $\boldsymbol{\eta}$ given $\boldsymbol{\varphi}_i$. Then the integrated likelihood corresponding to the observation θ_i is

$$L(\boldsymbol{\varphi}_i \mid \theta_i) = \int_{\boldsymbol{\eta}} f(\theta_i; \boldsymbol{\varphi}_i, \boldsymbol{\eta}) \pi(\boldsymbol{\eta} \mid \boldsymbol{\varphi}_i) d\boldsymbol{\eta} .$$

The joint likelihood for the n observations is then formed as the product of the n integrated likelihoods for single observations, that is,

$$L(\boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_n \mid \theta_1, \cdots, \theta_n) = \prod_{i=1}^{n} L(\boldsymbol{\varphi}_i \mid \theta_i)$$

We refer to the above *L* as "likelihood integrated". In what follows we use the likelihood integrated to draw inferences about the location of change-point in a given set of angular data.

In the classical change-point problem set-up, we are interested to test the null hypothesis

$$H_0 : \boldsymbol{\varphi}_1 = \boldsymbol{\varphi}_2 = \cdots = \boldsymbol{\varphi}_n$$

against the alternative that

$$H_1 : \text{There exist r}, 1 \le \text{r} \le \text{n - 1 such that } \boldsymbol{\varphi}_1 = \cdots = \boldsymbol{\varphi}_r \ne \boldsymbol{\varphi}_{r+1} = \cdots \boldsymbol{\varphi}_n .$$

The estimate of the change-point is derived as a by-product of the testing procedure. Note that in the above we are testing for the presence of at most one change-point. When the change-point is at *r*, the likelihood integrated has the form

$$L(r, \boldsymbol{\varphi}_1, \boldsymbol{\varphi}_n \mid \theta_1, \cdots, \theta_n) = \prod_{i=1}^{r} L(\boldsymbol{\varphi}_1 \mid \theta_i) \prod_{i=r+1}^{n} L(\boldsymbol{\varphi}_n \mid \theta_i),\ 1 \le r \le n$$

where we define $\prod_{i=n+1}^{n} L(\boldsymbol{\varphi}_n \mid \theta_i) = 1$. It may be noted that when no change-point is present in the data set we have *r* = *n*. It may be noted that if $\pi(\boldsymbol{\eta} \mid \boldsymbol{\varphi}_i)$ is a proper prior then $L(\boldsymbol{\varphi}_i \mid \theta_i) = \int_{\boldsymbol{\eta}} f(\theta_i ; \boldsymbol{\varphi}_i, \boldsymbol{\eta}) \pi(\boldsymbol{\eta} \mid \boldsymbol{\varphi}_i) d\boldsymbol{\eta}$ is a proper density when viewed as a function of θ_i. Hence, it is simple to construct the generalized likelihood ratio (GLR) test for this problem using the likelihood integrated. The GLR test for testing H_0 against H_1 using likelihood integrated is

$$\Lambda = \frac{\sup_{\phi_1} L(n, \boldsymbol{\varphi}_1, \boldsymbol{\varphi}_n)}{\sup_{\phi_1, \phi_n, r} L(r, \boldsymbol{\varphi}_1, \boldsymbol{\varphi}_n)} < c$$

where *c* is chosen depending on the level of significance of the test. In this paper instead of the formal testing approach discussed above we focus on exploratory graphical analysis of the data with a possible change-point.

It is easy to see that the change-point problem can be viewed as a model selection problem. It can be thought of as a problem of choice among n models $M_r, 1 \le r \le n$ where M_r is the model with change-point at *r* for

$1 \le r \le n-1$ and M_n is the model with no change-point. The model selection may now be done on the basis of the likelihood integrated for the models M_r. Instead of a formal method, we suggest a graphical analysis useful for exploratory analysis. We find for each model M_r the value $L^*(r) = \sup_{\phi_1,\phi_n} L(r, \varphi_1, \varphi_n)$. The values of $L^*(r)$ are then plotted against *r*. It is seen that the plot of $L^*(r)$ shows a sharp downward trend beyond a point. This point can be thought of as rough estimate of the change point. This method of graphical examination is explained in more detail with the help of real life data sets later in the paper.

As a first illustration of the above approach, we discuss the change-point problem for the mean direction of the circular normal distribution. This problem has been studied in details from a frequentist perspective in Laha (2001). The models M_r for $1 \le r \le n-1$ are

$$\Theta_1,...,\Theta_r \textit{ are i.i.d. } CN(\mu_1,\kappa) \textit{ and } \Theta_{r+1},...,\Theta_n \textit{ are i.i.d. } CN(\mu_n,\kappa), \mu_n \ne \mu_1$$

and the model M_n is

$$\Theta_1, \Theta_2,...,\Theta_n \textit{ are i.i.d. } CN(\mu_1,\kappa).$$

We write $\mu_n = \mu_1 + \delta (\bmod 2\pi)$ where $0 \le \delta < 2\pi$. The parameters μ_1, δ, κ *and* r are assumed to be all unknown. The parameters of interest are (r, δ) and the nuisance parameters are (μ_1, κ). Henceforth we write μ_1 as μ. Let $\gamma(\mu, \kappa)$ be a proper joint density. The integrated likelihood for single observation is

$$L(\delta_i \mid \theta_i) = \int_0^\infty \int_0^{2\pi} \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta_i - \mu - \delta_i)} \gamma(\mu, \kappa) d\mu \, d\kappa$$

where $\delta_i = \begin{matrix} 0 & if\ 1 \le i \le r \\ \delta & if\ r+1 \le i \le n \end{matrix}$. Then the likelihood integrated for the n observations is $L(r,\delta \mid \theta_1,\cdots,\theta_n) = \prod_{i=1}^{r} L(0 \mid \theta_i) \prod_{i=r+1}^{n} L(\delta \mid \theta_i)$. We can then obtain $L^*(r) = \sup_{\delta} L(r,\delta \mid \theta_1,\cdots,\theta_n)$ and use the plot of $(r, L^*(r))$ for indication about possible presence of change-point.

As a second illustration, we look at the change-point problem for the parameter ν of Papakonstantinou's $P(\nu,k)$ distribution. We will assume that the initial value of ν is known and with no loss of generality, take it to be 0. Thus the models M_r for $1 \le r \le n-1$ are

$$\Theta_1,\ldots,\Theta_r \text{ are i.i.d. } P(0,k) \text{ and } \Theta_{r+1},\ldots,\Theta_n \text{ are i.i.d. } P(\nu_n,k), \nu_n \ne 0$$

and the model M_n is $\Theta_1,\Theta_2,\ldots,\Theta_n$ *are i.i.d.* $P(0,k)$. Let $\gamma(k)$ be a prior for *k*. Then the integrated likelihood for one observation is

$$L(\nu_i;\theta_i) = \frac{1}{2\pi} + \frac{E_\gamma(k)}{2\pi}\sin(\theta_i + \nu_i \sin\theta_i)$$

where $\nu_i = \begin{matrix} 0 & if\ 1 \le i \le r \\ \nu & if\ r+1 \le i \le n \end{matrix}$. It may be observed that $L(\nu_i;\theta_i)$, when viewed as a function of θ_i, is the density of $P(\nu_i, E_\gamma(k))$. The likelihood integrated for the n observations is $L(r,\nu \mid \theta_1,\cdots,\theta_n) = \prod_{i=1}^{r} L(0 \mid \theta_i) \prod_{i=r+1}^{n} L(\nu \mid \theta_i)$. We can then proceed as before to obtain $L^*(r) = \sup_{\nu} L(r,\nu \mid \theta_1,\cdots,\theta_n)$ and then use the plot of $(r, L^*(r))$ for indication about possible location of change-point.

Example

In this section we apply the above methodology for exploratory analysis of two real-life data sets – wind data set (Weijers et. al., 1995) and flare data set (Lombard, 1986).

We first discuss the exploratory analysis for the data set of wind directions given by Weijers et. al. (1995). They investigated the horizontal perturbation wind field within thermal structures encountered in the atmospheric surface layer boundary. We are interested to study the possible existence of change - point in the direction of the horizontal wind field. The Changeogram obtained by using DDSTAP1.1 (SenGupta, 2005) is given in Fig. 1 below. The Changeogram indicates the presence of two change-points. Since the method discussed in this paper is for detection of at most one change-point we consider only the first 22 observations for further analysis using the likelihood integrated method. We take the joint prior $\gamma(\mu,\kappa)=\dfrac{1}{12\pi}$ which arises as the product of two independent priors – a circular uniform prior on μ and a Uniform (0,6) prior on κ. The computations are done using the R software package. The plot of $\log L^*(r)$ against *r* is given in Fig. 2 below

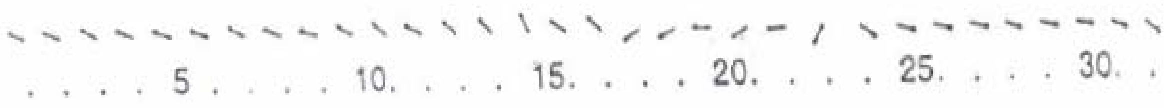

Fig 1: Changeogram of wind data

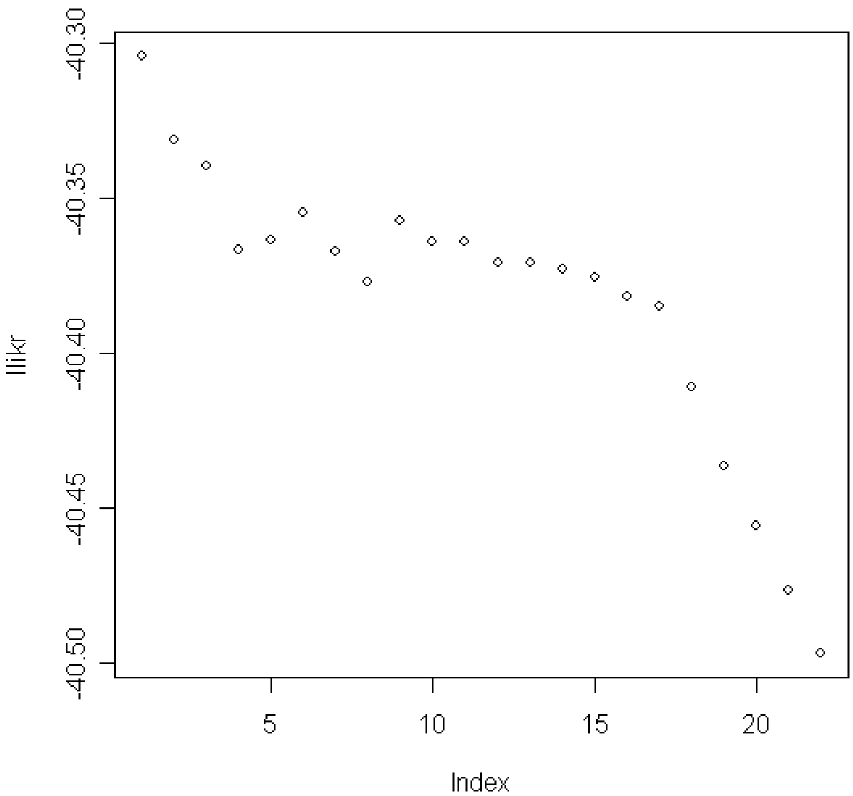

Fig 2: The plot of $\log L^*(r)$ against r for wind data

We find from the plot of $\log L^*(r)$ against r given in Fig. 2 that $\log L^*(r)$ declines very sharply after observation 17 which is the point of change as seen from the Fig. 1. The above dataset has been analyzed from a frequentist viewpoint in Laha (2001). It is reported therein that the NRTT, a parametric test derived under the assumption of circular normality in Laha (2001), when applied to this dataset indicates the presence of change-point at 5% level of significance and identifies the 17th observation as the change-point. When Lombard's non-parametric test (Lombard, 1986) for single change-point is applied to this dataset it indicates the presence of a change-point at 5% level of significance but identifies 13 as the change-point.

As a second example we analyze the flare data of Lombard (1986). The Changeogram is this data set is given in Fig. 3 below. The data set has been previously analyzed using a non-parametric framework in Lombard (1986) and using the NRTT in Laha (2001). Both these analyses indicated presence of more than one change-point in this data set. Based on the findings of the above mentioned studies, we consider two subsets of the full data set each consisting of at most one change-point. These subsets are chosen as (1) consisting of observation numbers 1-42 and (2) consisting of observation numbers 13-60. We use the same joint prior for (μ, κ) as in the case of wind data. The plots of $\log L^{*}(r)$ against *r* for the two subsets are given in Figs. 4 and 5 respectively.

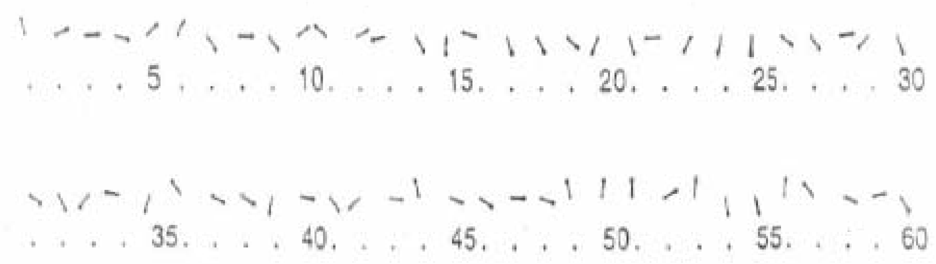

Fig 3. Changeogram for Flare data

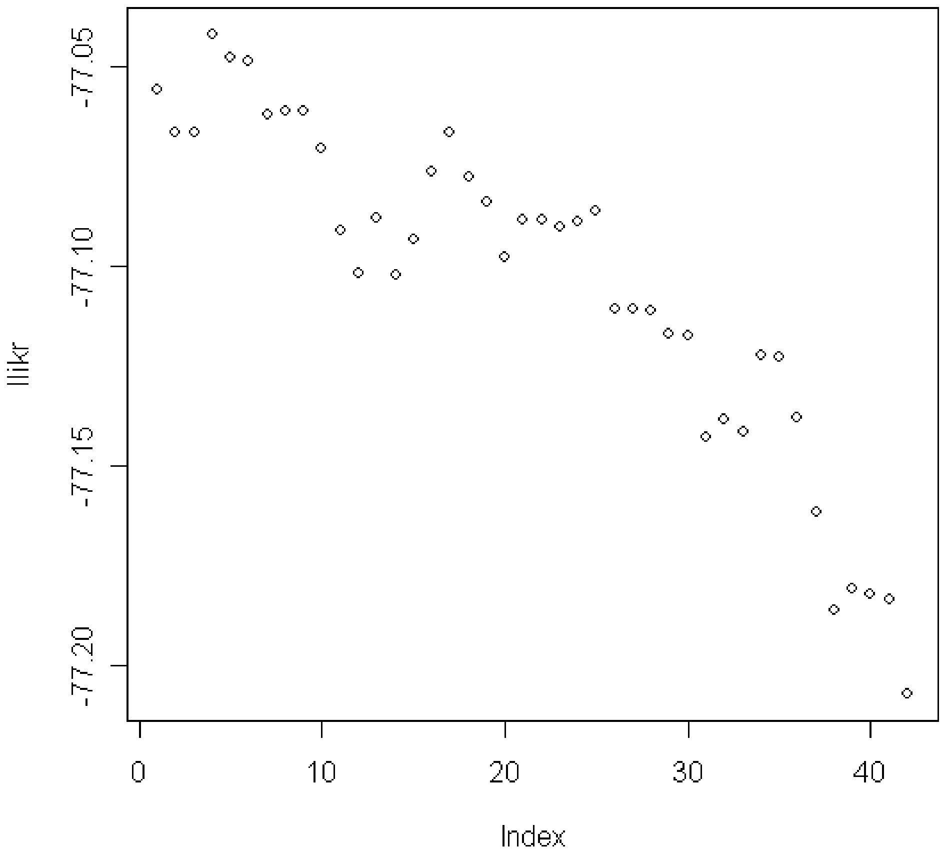

Fig 4: The plot of $\log L^{*}(r)$ against *r* for flare data (Obs. 1-42)

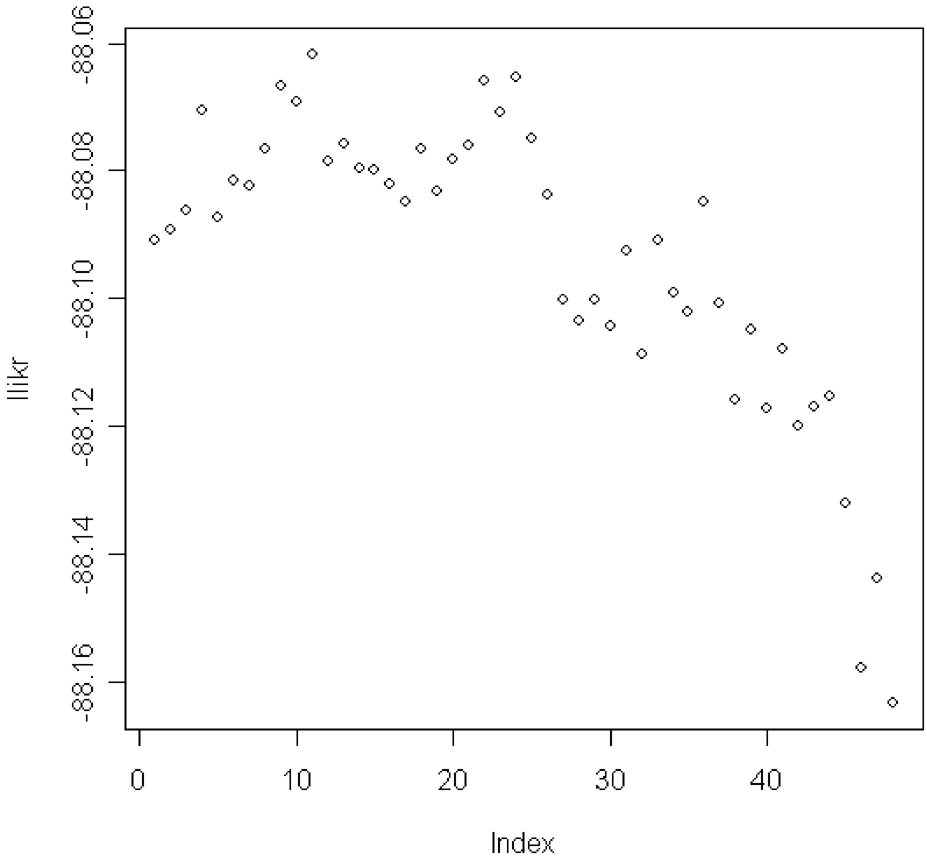

Fig 5: The plot of $\log L^{*}(r)$ against *r* for flare data (Obs. 13-60)

From Fig. 4 it is seen that there is a sharp decline in the values of $\log L^{*}(r)$ after observation number 25. A careful examination of Fig. 3 for this portion of the data set indicates the presence of a possible change-point at 25. However, frequentist analyses presented in Lombard (1986) and Laha (2001) did not indicate the presence of change-point at 25 and instead pointed to the observation number 12 as the change-point.

In Fig. 5 we see a sharp decline in the $\log L^{*}(r)$ values after the 26th observation in the data set 13-60, that is, after the 38th observation in the original data set. In this connection it may be noted that different tests applied to this data set had indicated different change-points in the neighborhood of the 38^{th} observation. Lombard's two change-point test (Lombard, 1986) indicated a presence of change-point at 36^{th} observation while Lombard's single change-point test and the NRTT (Laha, 2001) applied in the adaptive manner as in this paper indicated 42 as the change-point.

References:

1. Batschelet, E. (1981): *Circular Statistics in Biology*, Academic Press, London
2. Berger, J. O., Liseo, B. and Wolpert, R. L. (1999): Integrated likelihood methods for eliminating nuisance parameters, *Statistical Science*, 14,1,1-22.
3. Chen, J. and Gupta, A. K. (2000): *Parametric Statistical Change Point Analysis*, Birkhauser, Boston.
4. Chernoff, H. and Zacks, S. (1964) : Estimating the current mean of a normal distribution which is subject to changes in time, *Annals of Mathematical Statistics*, 35, 999-1018.
5. Jammalamadaka, S. R and SenGupta, A (2001): *Topics in Circular Statistics*, World Scientific, Singapore
6. Laha, A. K. (2001): *Slippage and Change-point Problems with Directional Data*, Ph.D. thesis, Indian Statistical Institute, Kolkata, India.
7. Lombard, F. (1986): The change-point problem for angular data : A nonparametric approach, *Technometrics,* 28, 391-397.

8. Page, E. S. (1955) : A test for a change in a parameter occurring at an unknown time point, *Biometrika*, 42, 523-526.
9. Sen, A. K. and Srivastava, M. S. (1973): On multivariate tests for detecting change in mean, *Sankhya*, A, 35, 173-185.
10. Sen, A. K. and Srivastava, M. S. (1975a): On tests for detecting a change of mean, *Annals of Statistics*, 3, 98-108.
11. Sen, A. K. and Srivastava, M. S. (1975b): Some one-sided tests for change in level, *Technometrics*, 17, 61-64.
12. SenGupta, A. (2005). DDSTAP 1.1 – Directional Data STatistical Analysis Package 1.1. Indian Statistical Institute, Kolkata.
13. Weijers, E.P., Van Delden, A., Vugts, H.F., and Meesters, A.G.C.A. (1995): The composite horizontal wind field within convective structures of the atmospheric surface layer, *Journal of the Atmospheric Sciences*, 52, 3866-3878.