Gazi University Journal of Science
GU J Sci
26(4):543-555 (2013)

Investigation of the Change Point in Mean of Normal Sequence Having an Outlier

Ayten YİĞİTER [1,♠], Meral ÇETİN [2]

[1] *Hacettepe University, Statistics Department, Beytepe Ankara*
[2] *Hacettepe University, Statistics Department, Beytepe, Ankara*

Received: 07.10.2013 Accepted:03.11.2013

ABSTRACT
In this study, the change point in mean of the sequence of the random variables from normal distribution under the case of having an outlier in the sequence is considered. Under with this case, the maximum likelihood estimate of the change point and the estimates of the change point using robust methods are computed. The performances of the maximum likelihood method and robust methods on the estimation of the change point according to outlier locations with different sample sizes are investigated via extensive simulation studies.

Key words: Change point; outlier; maximum likelihood method; robust methods; simulation.

1. INTRODUCTION

Let X_i, $i=1,\ldots,n$ comes from normal distribution with different mean μ_i and variance σ^2. In this sequence, single change point model is defined as follows:

$$X_i \sim N(\mu_1,\sigma^2), \qquad i=1,\ldots,\tau$$
$$X_i \sim N(\mu_2,\sigma^2), \qquad i=\tau+1,\ldots,n\ . \qquad (1)$$

Here τ is an unknown change point, μ_1, μ_2 are unknown means and σ^2 is known common variance. The case defined in Eq.(1) is a well-known problem called change point problem in the literature. Maximum likelihood approach is commonly used to estimate the change point as well as some other techniques. As it is known in the literature, maximum likelihood estimate of the change point could be affected by outliers. Therefore, robust estimation techniques should be considered to minimize the effect of an outlier.

The change point problem has been studied since 1950s (Jarrett, 1979). There are many studies on the estimating and the hypothesis testing of one change point problem in normal, binomial, Poisson, exponential and gamma distributed random variables. For the normal distributed random variables, Hinkley (1970), and Hinkley and Hinkley (1970), Worsley (1987), Chen and Gupta (1997, 2004) investigated the change point in the sequence of normal distributed random variables. Some of the fundamental works on the change point in the sequence of exponential and Poisson distributed random variables are available in Haccou et al. (1988); Jandhyala and Fotopoulos (1999, 2001); Boudjelaba, et al. (2001). Fotopoulos and Jandhyala (2001). Ramanayake (2004) considers tests for a change in the

♠Corresponding author, e-mail:yigiter@hacettepe.edu.tr

shape parameter of gamma distributed random variables.

Even though there have been many studies concerning the change point problem in the literature, there are a few studies dealing with an outlier. For example, Takeuchi and Yamanishi (2006) are interested in the issues of outlier detection and change point detection from time series. Pechenizkiy et al. (2009) gave a framework based on switching regression models depending on perceived changes in the data from a pilot circulating fluidized bed reactor.

Outliers can be present in the data for many different reasons. Influence of outliers on the parameter estimation is investigated. Especially in regression analysis where outliers have bad effects, robust methods are developed. Huber (1973) introduced the class of M estimators as a robust technique. Alternative robust estimators have been developed by Hinich and Talwar (1975) and Andrews (1974). Harvey (1977) compared the robust methods which are related to median. O'leary (1990) compared the four weight functions by used reweighted least squares.
In this study, we focus on the estimation of change point in the mean of the sequence normal random variables having an outlier. To estimate the change point, we generate some data from the model given Eq.(1) and we add a single outlier as a fixed point in the sequence and we showed the influence of the outlier on the maximum likelihood estimator of the change point. Since robust estimation techniques are considered as alternative in order to decrease the influence of the outlier, we investigate the robust methods for the estimate of the change point by extensive simulation study. Results are compared according to the mean squared error (*mse*) and the relative frequencies of the estimates of change point being equal to the true change point location.

2. MAXIMUM LIKELIHOOD ESTIMATOR OF CHANGE POINT

Under the change point model given Eq.(1), the likelihood function for observed random variables $x_1,\ldots,x_n$ is:

$$L(\mu_1,\mu_2,\sigma^2,\tau|x_1,x_2,\ldots,x_n)=L(\mu_1,\mu_2,\sigma^2,\tau)=\frac{1}{2\pi\sigma^2}e^{-\sum_{i=1}^{\tau}\frac{(x_i-\mu_1)^2}{2\sigma^2}-\sum_{i=\tau+1}^{n}\frac{(x_i-\mu_2)^2}{2\sigma^2}}.$$

The log (ℓn) likelihood function can be written as:

$$\ell nL(\mu_1,\mu_2,\sigma^2,\tau)=-\ell n(2\pi\sigma^2)-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\tau}(x_i-\mu_1)^2+\sum_{i=\tau+1}^{n}(x_i-\mu_2)^2\right\}. \quad (2)$$

For given a fixed value of τ, the maximum estimators of μ_1 and μ_2 are obtained

$$\hat{\mu}_1=\frac{\sum_{i=1}^{\tau}x_i}{\tau}=\bar{x}_1, \qquad \hat{\mu}_2=\frac{\sum_{i=\tau+1}^{n}x_i}{n-\tau}=\bar{x}_2$$

and these estimates in Eq.(2) replace μ_1 and μ_2, Eq.(2) can be written as:

$$\ell nL(\sigma^2,\tau)=-\ell n(2\pi\sigma^2)-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\tau}(x_i-\bar{x}_1)^2+\sum_{i=\tau+1}^{n}(x_i-\bar{x}_2)^2\right\}. \quad (3)$$

Eq.(3) is the function of τ only for known the variance, σ^2. The log likelihood in Eq.(3) is:

$$\ell nL(\tau)\propto-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\tau}(x_i-\bar{x}_1)^2+\sum_{i=\tau+1}^{n}(x_i-\bar{x}_2)^2\right\}. \quad (4)$$

Maximum likelihood estimate of τ is obtained by maximizing the function given in Eq.(4) (Hinkley,1970):

$$\hat{\boldsymbol{\tau}}=\arg\max_{\tau}\ell nL(\tau), \qquad \tau=1,2,\ldots,n-1$$

$$\hat{\boldsymbol{\tau}}=\arg\max_{\tau}\left\{-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\tau}(X_i-\bar{X}_1)^2+\sum_{i=\tau+1}^{n}(X_i-\bar{X}_2)^2\right\}\right\}, \quad \tau=1,2,\ldots,n-1. \quad (5)$$

3. ROBUST ESTIMATORS

Robust regression estimators were first suggested by Huber (1973) as M estimators in regression. Robust methods are based on the idea of minimizing another function of the residuals instead of minimizing the sum of squared residuals. Therefore various functions called influence or weight functions are proposed for residuals.

A regression model is:

$$y_i=\beta_0+\beta_1x_i+e_i, \qquad i=1,2,\ldots,n.$$

Here x_i and y_i are the predictor and response variable values, respectively, and e_i are random errors. Let $e_i=y_i-\hat{y}_i$ be residual of the i^{th} datum, the difference between the *i.* observation and its fitted value. The purpose of the M estimators is based on following minimization problem

$$\min_{\hat{\beta}}\sum_{i=1}^{n}\rho(e_i)$$

where ρ is a symmetric and positive-definite function and a unique minimum at zero. After the derivative of the function $\rho(.)$ with respect to β_j, we have

$$\sum_{i=1}^{n}\psi(e_i)x_i=0 \quad (6)$$

where $\psi(.)$ is the first derivative of the function $\rho(.)$ and it is called $\psi(.)$ an estimating function or just call it a ψ function.

In most cases, a solution for Eq.(6) can be found iteratively. Many functions are suggested in literature. Some of them are given Table 1 (Rousseeuw and Leroy, 1987, Bhar, 2011).

Table 1. Some commonly used M estimators

type	$\psi(x)$	*k*
Andrews	$\psi(x)=\begin{cases}\sin\left(\frac{x}{k}\right), & \lvert x\rvert \le k\pi \\ 0, & \lvert x\rvert > k\pi\end{cases}$	1.5 or 2
bisquare	$\psi(x)=\begin{cases}x(1-(x/k)^2)^2 & \lvert x\rvert \le k\pi \\ 0 & \lvert x\rvert > k\pi\end{cases}$	5 or 6
Cauchy	$\psi(x)=\frac{x}{1+(x/k)^2}$	2.385
Fair	$\psi(x)=x(1+\lvert x\rvert/k)^{-1}$	1.4
Huber	$\psi(x)=\begin{cases}-k, & x<-k \\ x, & -k\le x\le k \\ k, & x>k\end{cases}$	1.345
logistic	$\psi(x)=k\tanh(x/k)$	1.205
Talwar	$\psi(x)=\begin{cases}x, & \lvert x\rvert \le k \\ 0, & \lvert x\rvert > k\end{cases}$	2.985
Welsch	$\psi(x)=xe^{-(x/k)^2}$	2.985

4. PROPOSED METHOD

The proposed idea estimates the location of change point in the contaminated sequence by replacing μ_1 and μ_2 with robust estimates in Eq.(2) instead of using $\bar{X}_1$ and $\bar{X}_2$ which are known to have 0% breakdown point that is are influenced by an outlier in the sequence. For the robust estimates of μ_1 and μ_2, we used the M estimators, also the mode and the median in the simulation study.

Another alternative approach could be using *MAD*(median absolute deviation) to estimate the location of the change point. *MAD* is known to be a robust estimator of standard deviation(σ) and it is given as follows:

$$MAD = median\{|X_i - M|\}, \qquad i = 1,2,...,n.$$

Here M denotes the median of the sample.

In change point problem, we consider two *MAD*s with respect to change point in the sequence:

$$MAD_1 = median\{|X_i - M_1|\}, \qquad i = 1,2,...,\tau$$

$$MAD_2 = median\{|X_i - M_2|\}, \qquad i = \tau+1,\tau+2,...,n$$

where M_1 and M_2 are the medians of the sequence before and after the change point respectively. The expressions, $\sum_{i=1}^{\tau}(X_i-\bar{X}_1)^2$ and $\sum_{i=\tau+1}^{n}(X_i-\bar{X}_2)^2$, in Eq.(5) can be replaced to $(\tau-1)MAD_1^2$ and $(n-\tau-1)MAD_2^2$ respectively.

Our goal is to investigate the efficiency of suggested modifications in Eq.(5) and compare them via simulation studies.

5. SIMULATION STUDY

To evaluate the performances of the modifications explained in Section 4, random samples from normal distributions with sample sizes n=40, 60, 100 are generated and the change point location is fixed at the

first quartile, at the center and at the third quartile of the samples respectively.

For each sample, the values of parameters are taken arbitrarily as (μ_1 , μ_2) = (0.1, 0.3), (0.5, 1.5), (2.5, 7.5), (5, 15), (0.3, 0.1), (1.5, 0.5), (7.5, 2.5), (15, 5) for each standard deviations σ=0.001, 0.01, 0.1, 1, 1.5, 2, 2.5.

The number of iterations is taken to be 500 times for each sample size. The outliers are generated from a normal distribution with eight times the mean of the random samples. The outlier locations in the sequence are placed before the true change point location or after the true change point location as given in Table 2

Table 2. Outlier locations with respect to the true change point location.

Change point location	Outlier location with respect to the true change point location
25th % observation	20th % observation or 30th % observation
50th % observation	45th % observation or 55th % observation
75th % observation	70th % observation or 80th % observation

For example, let true change point location be 50th observation of the sequence from the sample of size n=100. The location of the outlier may be 45th observation of the sequence or 55th observation of the sequence (see Figure 1).

We computed the mean of estimates $\hat{\tau}$, $\overline{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated mean squared errors (*mses*) for each sample size and the value of parameters.

The scatter plot or the box plot of the samples could be used to investigate whether there is an outlier in the sequence, or not (see Figure 1). In addition, Grubbs's test can be used to detect the outlier in sequence. The null and alternative hypotheses for the test are defined by

H_0: There is no outlier in the sequence

H_1: There is at least one outlier in the sequence.

The test statistic of the Grubbs is,

$$G = \frac{\max\left|X_i - \overline{X}\right|}{S}$$

where S is the standard deviation of the sample. At significance level α=0.05, the test rejects H_0 if the test statistics, G, is greater than the critical value, $\frac{n-1}{n}\sqrt{\frac{t^2_{\alpha/2n,n-2}}{n-2+t^2_{\alpha/2n,n-2}}}$.

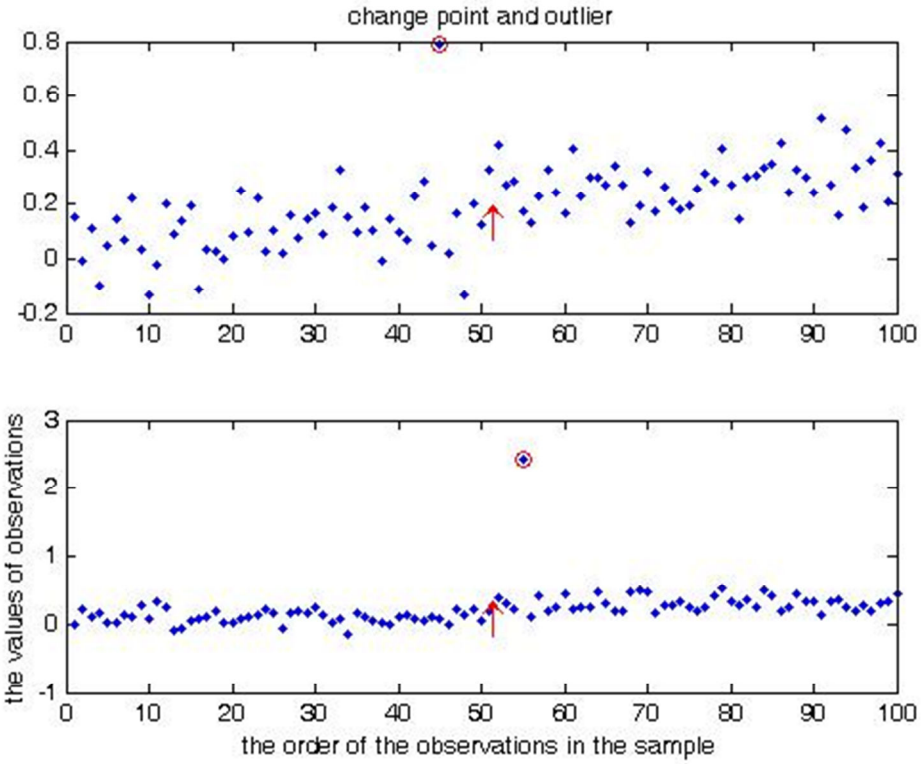

Figure 1. A change point and an outlier in the sequence with the sample size 100 and the parameters of normal distribution (μ_1 , μ_2)=(0.5, 1.5) and σ=0.1.

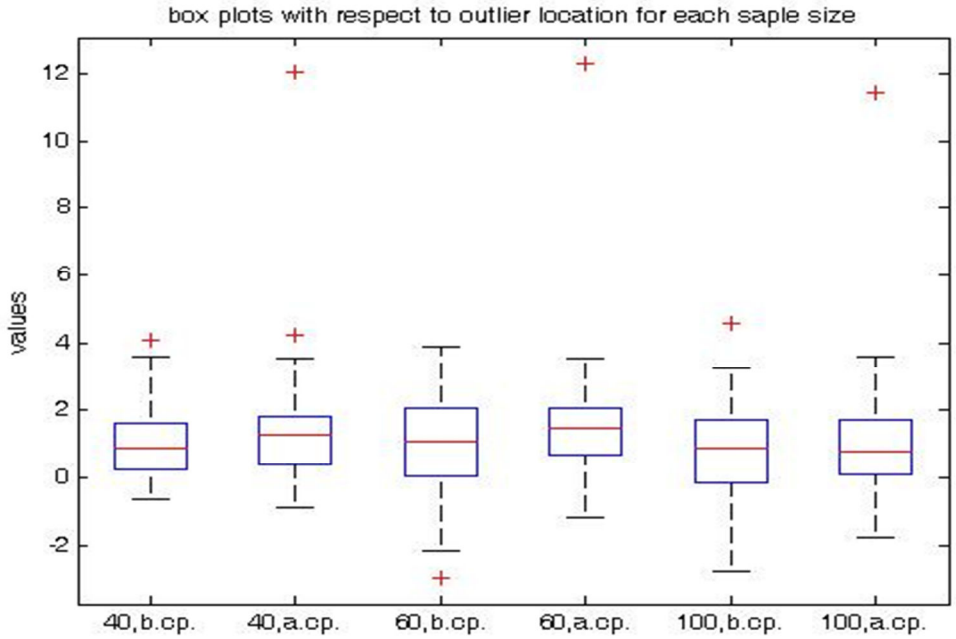

Figure 2. The box plot of the samples from normal distribution with parameters (μ_1 , μ_2)=(0.5, 1.5) and σ=0.001 and the change point being at the center of the sequence and the outlier with respect to change point location (b.cp: before the change point, a.cp: after the change point).

In the simulation study, when we test whether there is an outlier using the Grubbs's test, at each value of the means (μ_1 , μ_2) = (0.1, 0.3), (0.3, 0.1) for each standard deviation σ=1, 1.5, 2, 2.5 and at (μ_1 , μ_2) = (0.1, 0.3), (0.3, 0.1), (0.5, 1.5), (1.5, 0.5) for σ=2.5, the null hypothesis could not be rejected with the significance level α=0.05. The results are evaluated with in the case of being an outlier in the sequence.

6. RESULTS

The performances of maximum likelihood estimator of change point show some differences in results for taking standard deviation $\sigma < 0.1$ and $\sigma \geq 0.1$. That's why; the results are evaluated for these cases. Table 3 and Table 4 show the results for σ= 0.01 and (μ_1 , μ_2)= (0.1,0.3),(0.3,0.1) respectively. Also, the sample size n=100, true change point location (τ being 50th percentile observation) and outlier locations (being 45th percentile or 55th percentile) are shown by Table 3 and 4. From these tables, it is clear that robust estimates of μ_1 and μ_2, except the mode, did not provide the desired results. On the contrary to our expectation, robust estimators in Eq.(5) are not sufficient to estimate the true change point location when an outlier existing in the sequence. On the positive side, the sensitivity of the arithmetic mean to an outlier is clearly demonstrated.

Table 3. The mean of estimates $\hat{\tau}$, $\overline{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated *mse*s for n=100, τ=50th % observation and (μ_1 , μ_2) = (0.1, 0.3), σ=0.01.

	μ_1=0.10, μ_2=0.30, σ=0.01, n=100, τ=50					
	outlier =45th observation			outlier =55th observation		
estimators	$\overline{\hat{\tau}}$	*f*	*mse*	$\overline{\hat{\tau}}$	*f*	*mse*
mean	44.00	0.00	36.00	50.00	500.00	0.00
mode	50.06	484.00	0.16	50.06	486.00	0.19
median	3.69	0.00	2146.63	4.28	0.00	2093.37
Andrews	28.11	0.00	479.42	27.66	0.00	499.32
bisquare	28.12	0.00	479.15	27.63	0.00	500.75
Cauchy	26.76	0.00	540.63	28.35	0.00	469.49
Fair	34.05	0.00	255.01	35.26	0.00	217.80
Huber	31.33	0.00	349.52	32.75	0.00	298.69
logistic	31.26	0.00	351.80	32.60	0.00	303.74
Talwar	34.89	0.00	228.53	27.86	0.00	490.55
Welsch	97.00	0.00	2209.00	97.00	0.00	2209.00

Table 4. The mean of estimates $\hat{\tau}$, $\bar{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated *mse*s for n=100, τ=50th % observation and (μ_1 , μ_2) = (0.3, 0.1), σ=0.01.

	μ_1=0.30, μ_2=0.10, σ=0.01, n=100, τ=50					
	outlier =45th observation			outlier =55th observation		
estimators	$\bar{\hat{\tau}}$	*f*	*mse*	$\bar{\hat{\tau}}$	*f*	*mse*
mean	50.00	500.00	0.00	55.00	0.00	25.00
mode	49.96	488.00	0.11	49.97	491.00	0.08
median	95.78	0.00	2099.42	96.32	0.00	2147.27
Andrews	72.37	0.00	500.68	71.91	0.00	480.33
bisquare	72.39	0.00	501.84	71.91	0.00	480.15
Cauchy	71.68	0.00	470.41	73.26	0.00	541.48
Fair	64.76	0.00	218.33	66.05	0.00	257.91
Huber	67.29	0.00	300.09	68.72	0.00	351.36
logistic	67.42	0.00	304.48	68.83	0.00	355.51
Talwar	72.13	0.00	490.17	65.12	0.00	228.87
Welsch	14.71	7.00	1503.49	14.30	8.00	1530.67

In case of $\sigma \geq 0.1$, Tables 5-8 show the result for (μ_1 , μ_2)={(0.1,0.3), (0.5,1.5), (2.5,7.5), (5,15)} σ = 0.1, sample size n=100, true change point (τ is 50th percentile), and the outlier locations (45th and 55th percentiles) in the sequence.

From Table 5, it is seen that all robust estimates and the arithmetic mean are insufficient to estimate the true change point location, when the difference between μ_1 and μ_2 is small. But, this situation changes gradually for the arithmetic mean and the mode, as the difference between μ_1 and μ_2 increase (See Tables 6-8). We focus on especially the arithmetic mean and the mode in the following of the study.

Table 5. The mean of estimates $\hat{\tau}$, $\bar{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated *mse*s for n=100, τ=50th % observation and (μ_1 , μ_2) = (0.1, 0.3), σ=0.1.

	μ_1=0.10, μ_2=0.30, σ=0.10, n=100, τ=50					
	outlier =45th observation			outlier =55th observation		
estimators	$\bar{\hat{\tau}}$	*f*	*mse*	$\bar{\hat{\tau}}$	*f*	*mse*
mean	46.11	115.00	27.85	50.33	300.00	1.54
mode	81.22	15.00	1407.30	82.10	18.00	1442.99
median	37.49	5.00	315.90	40.74	12.00	284.27
Andrews	45.83	69.00	34.29	48.63	150.00	11.74
bisquare	45.75	68.00	34.73	48.65	146.00	11.66
Cauchy	44.07	19.00	43.41	49.23	172.00	7.35
Fair	43.98	3.00	44.94	49.81	230.00	5.35
Huber	44.27	18.00	42.69	49.27	163.00	8.00
logistic	43.96	8.00	45.21	49.55	192.00	5.95
Talwar	46.39	99.00	35.71	48.01	128.00	16.03
Welsch	91.06	0.00	2174.26	90.50	0.00	2178.60

Table 6. The mean of estimates $\hat{\tau}$, $\bar{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated *mse*s for n=100, τ=50th % observation and (μ_1 , μ_2) = (0.5, 1.5), σ=0.1.

	μ_1=0.50, μ_2=1.50, σ=0.10, n=100, τ=50					
	outlier=45th observation			outlier=55th observation		
estimators	$\bar{\hat{\tau}}$	*f*	*mse*	$\bar{\hat{\tau}}$	*f*	*mse*
mean	44.14	12.00	35.14	50.00	500.00	0.00
mode	50.19	473.00	0.91	50.13	477.00	0.54
median	6.91	0.00	1869.08	7.99	0.00	1779.91
Andrews	30.21	0.00	392.46	30.01	0.00	400.69
bisquare	30.21	0.00	392.70	29.99	0.00	401.18
Cauchy	31.92	0.00	330.40	33.55	0.00	274.06
Fair	35.40	0.00	214.00	36.56	0.00	181.59
Huber	33.61	0.00	270.33	34.86	0.00	230.99
logistic	33.89	0.00	260.90	35.23	0.00	219.58
Talwar	35.36	0.00	215.01	29.83	0.00	407.75
Welsch	97.00	0.00	2209.00	97.00	0.00	2209.00

Table 7. The mean of estimates $\hat{\tau}$, $\bar{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated *mse*s for n=100, τ=50th % observation and (μ_1 , μ_2) = (2.5, 7.5), σ=0.1.

	μ_1=2.50, μ_2=7.50, σ=0.10, n=100, τ=50					
	outlier =45th observation			outlier =55th observation		
estimators	$\bar{\hat{\tau}}$	f	mse	$\bar{\hat{\tau}}$	f	mse
mean	44.00	0.00	36.00	50.00	500.00	0.00
mode	50.00	498.00	0.00	50.00	498.00	0.00
median	3.02	0.00	2206.79	3.09	0.00	2201.10
Andrews	26.62	0.00	546.67	25.99	0.00	576.51
bisquare	26.65	0.00	545.54	25.97	0.00	577.58
Cauchy	24.45	0.00	653.16	26.05	0.00	573.48
Fair	33.05	0.00	287.41	34.15	0.00	251.35
Huber	28.95	0.00	443.73	30.42	0.00	383.79
logistic	28.34	0.00	469.64	29.82	0.00	407.52
Talwar	34.25	0.00	248.31	26.39	0.00	557.67
Welsch	97.00	0.00	2209.00	97.00	0.00	2209.00

Table 8. The mean of estimates $\hat{\tau}$, $\bar{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated *mse*s for n=100, τ=50th % observation and (μ_1 , μ_2) = (5, 15), σ=0.1.

	μ_1=5.00, μ_2=15.00, σ=0.10, n=100, τ=50					
	outlier =45th observation			outlier =55th observation		
estimators	$\bar{\hat{\tau}}$	f	mse	$\bar{\hat{\tau}}$	f	mse
mean	44.00	0.00	36.00	50.00	500.00	0.00
mode	50.00	500.00	0.00	50.00	500.00	0.00
median	3.00	0.00	2209.00	3.00	0.00	2209.00
Andrews	26.00	0.00	576.00	25.08	0.00	620.88
bisquare	26.00	0.00	576.00	25.05	0.00	622.55
Cauchy	23.98	0.00	677.27	25.13	0.00	618.83
Fair	32.75	0.00	297.89	34.00	0.00	256.14
Huber	27.39	0.00	511.31	28.86	0.00	447.31
logistic	26.83	0.00	536.94	28.21	0.00	475.12
Talwar	34.00	0.00	255.94	26.00	0.00	576.00
Welsch	97.00	0.00	2209.00	97.00	0.00	2209.00

It is know that the arithmetic mean is influenced by outliers in data. But in the case of the change point, it is interesting to see that the influence of the outlier depends on its location where it appears in the sequence. The conclusion is undesirable when the first aim of study is to detect the change point in the sequence. When we look at the performance of the mode, the mode is consistent regardless of the outlier location in the sequence (the left of Figure 3). From Figure 3, the change point is not detected by the arithmetic mean when the outlier occurred before the change point for (μ_1 , μ_2) = (0.1, 0.3), σ=0.01, in contrast to the relative frequencies of the estimate $\hat{\tau}$ being equal to the true change point is very high for using the mode. These can be seen from the estimated *mses* of the mode for each sample size.

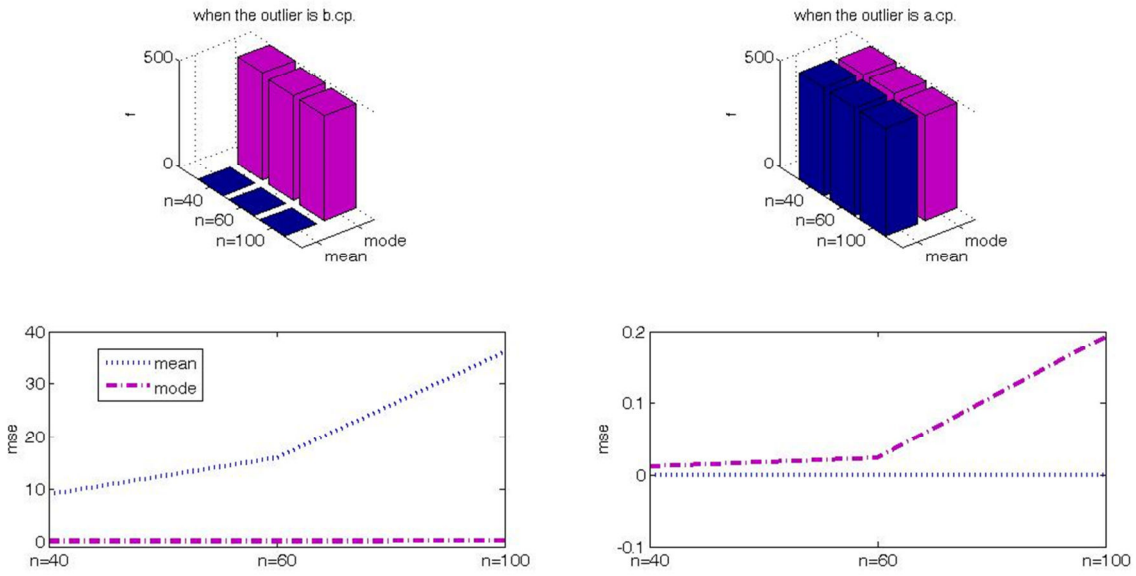

Figure 3. The mean of estimates $\hat{\tau}$, $\bar{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated *mses* for the arithmetic mean and the mode with respect to outlier location for the sample sizes n=40, 60, 100, (μ_1 , μ_2)=(5, 15), σ=0.1 and τ=50th % observation.

When the outlier occurred after the change point for (μ_1 , μ_2)=(0.1, 0.3), σ=0.01, both the arithmetic mean and the mode have a good performance to detect to the change point in the sequence. Despite the estimated *mse*s are close to zero for both, the estimated *mses* increase in the case of using the mode as the sample size increases (the right of Figure 3).

It is seen that when using the arithmetic mean and the mode, the relative frequencies of the estimate $\hat{\tau}$ which equals to the true change point are almost the same, the estimated *mse* is higher in the case of using the mode than the estimated *mse* in the case of using the arithmetic mean. It means that when the true change point is not detected, the estimation of the change point for using the mode is found to close to the end of the sequence and so the estimated *mse* increases in the case of using the mode.

When the variances are greater than 0.1, especially, as the difference between μ_1 and μ_2 are low, the arithmetic mean or the mode is not adequate for the estimate of the change point even regardless of the location of the outlier. The arithmetic mean has either a good performance or bad performance to estimate the change point with respect to the outlier location in the sequence. But the behavior of the mode is surprisingly different, for example, as (μ_1 , μ_2)=(0.1, 0.3), (0.3, 0.1) and σ=0.1, it has the biggest estimated *mse*. Despite that, at the same variance, as the difference between μ_1 and μ_2 increases, it is seen that the estimated *mses* sharply decrease and the relative frequencies of the estimate $\hat{\tau}$ being equal to the true change point increase (Figure 4). This case cannot be explained directly the amount of the variance or the difference between μ_1 and μ_2 . Especially, this behavior of the mode could be explained as follows: as known, about 99.99% of the observations fall within 4 standard deviations of the mean for the normal distributed samples. In the case of having the change point in the sequence, there are two independent sample from normal distribution defined Eq.(1) such that about

99.99% of the observations will fall in the interval (μ_1 -4σ, μ_1 +4σ) before the change point, and about 99.99% of the observations will fall in the interval (μ_2 -4σ, μ_2 +4σ) after the change point. If there is an intersection between these intervals, this is an important factor to estimate the change point in the case of using the mode. Also this factor is valid in the case of using the arithmetic mean, but the mode is more sensitive to be how much of the intersection is between the intervals (μ_1 -4σ, μ_1 +4σ) and (μ_2 -4σ, μ_2 +4σ) (see Table 9).

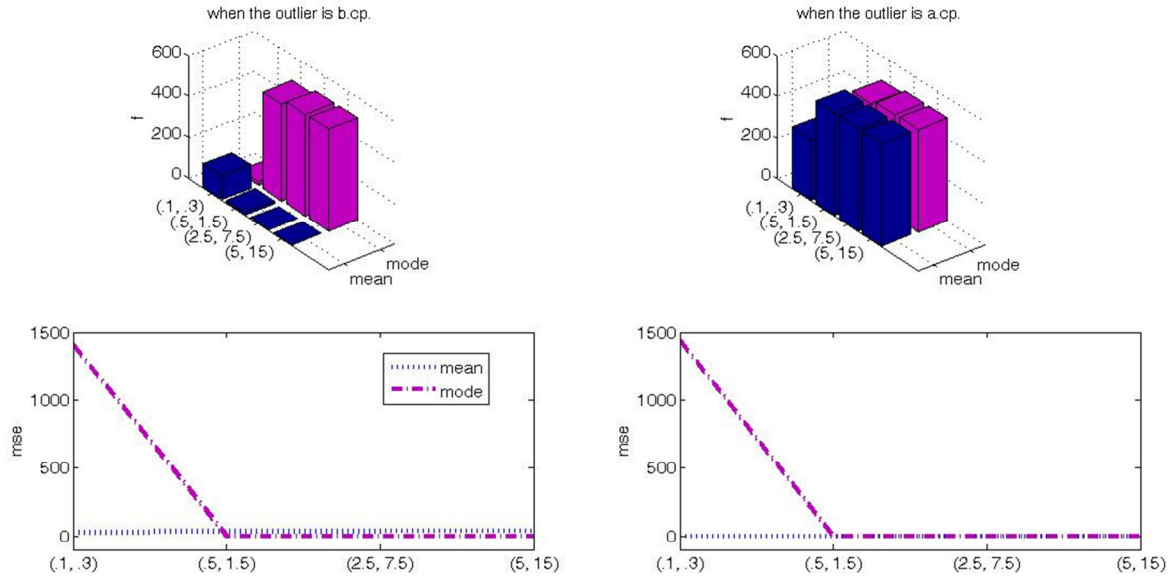

Figure 4. The mean of estimates $\hat{\tau}$, $\bar{\hat{\tau}}$, and the relative frequencies (f) of the estimate $\hat{\tau}$ being equal to the true change point and the estimated *mse*s for the arithmetic mean and the mode with respect to the outlier location and (μ_1 , μ_2) = (0.1, 0.3), (0.5, 1.5), (2.5, 7.5), (5, 15), and σ=0.1 for the sample size n=100 and τ=50th % observation.

Table 9. The range of the observations from normal distribution with σ=0.1 and (μ_1 , μ_2)=(0.1, 0.3), (0.5, 1.5), (2.5, 7.5), (5, 15).

(μ_1 , μ_2)	The sample before the change point		The sample after the change point	
	μ_1 –4σ	μ_1 +4σ	μ_2 –4σ	μ_2 +4σ
(0.1, 0.3)	-0.30	0.50	-0.10	0.70
(0.5, 1.5)	0.10	0.90	1.10	1.90
(2.5, 7.5)	2.10	2.90	7.10	7.90
(5, 15)	4.60	5.40	14.60	15.40

When we use the *MAD*s with replace to $\sum_{i=1}^{\tau}\left(X_i-\bar{X}_1\right)^2$ and $\sum_{i=\tau+1}^{n}\left(X_i-\bar{X}_2\right)^2$, results are given in Table 10-11 respectively for the cases $\sigma < 0.1$ and $\sigma \geq 0.1$.

Table 10. The results of the mean and the *MADs* for n=100, τ=50th % observation and σ=0.01.

		σ=0.01, n=100, τ=50					
		outlier=45th observation			outlier=55th observation		
(μ_1, μ_2)	**estimators**	$\bar{\hat{\tau}}$	*f*	*mse*	$\bar{\hat{\tau}}$	*f*	*mse*
(0.10 ,0.30)	*mean*	44.00	0.00	36.00	50.00	500.00	0.00
	MADs	49.63	139.00	8.61	50.14	155.00	9.71
(0.5, 1.5)	*mean*	44.00	0.00	36.00	50.00	500.00	0.00
	MADs	49.57	153.00	10.45	50.12	156.00	8.05
(2.5, 7.5)	*mean*	44.00	0.00	36.00	50.00	500.00	0.00
	MADs	49.72	142.00	9.47	50.13	153.00	9.38
(5, 15)	*mean*	44.00	0.00	36.00	50.00	500.00	0.00
	MADs	49.86	160.00	7.47	50.14	143.00	10.44

Table 11. The results of the mean and the *MADs* for n=100, τ=50th % observation and σ=0.1.

		σ=0.1, n=100, τ=50					
		outlier=45th observation			outlier=55th observation		
(μ_1, μ_2)	**estimators**	$\bar{\hat{\tau}}$	*f*	*mse*	$\bar{\hat{\tau}}$	*f*	*mse*
(0.10 ,0.30)	*mean*	45.92	95.00	28.27	50.30	301.00	1.37
	MADs	50.03	98.00	56.65	50.53	97.00	54.30
(0.5, 1.5)	*mean*	44.12	10.00	35.28	50.00	500.00	0.00
	MADs	49.75	153.00	8.17	50.17	144.00	10.11
(2.5, 7.5)	*mean*	44.00	0.00	36.00	50.00	500.00	0.00
	MADs	49.54	159.00	9.11	50.11	146.00	10.31
(5, 15)	*mean*	44.00	0.00	36.00	50.00	500.00	0.00
	MADs	49.80	151.00	8.70	50.11	144.00	10.14

From Tables 10-11, we see that is no advantage of using *MAD* to reduce the influence of the outlier in the sequence.

7. CONCLUDING REMARKS

In this study, we investigated the influence of the outlier on the estimation of the change point in normal distributed sequence via a

simulation study. According to the simulation results, we can see that robust estimators given in Section 3 are insufficient to estimate of the change point.

The influence of the outlier on the estimation of the change point in the sequence is reduced by using the mode in Eq.(5), as to be one of the conclusion, it is seen that the robust methods are necessary for the change point problem in the sequence with an outlier.

REFERENCES

[1] Andrews, D. F., 1974. A Robust Method for Multiple Linear Regression. Technometrics, 16, 523-531.

[2] Bhar, L., "Robust Regression" web page: http://www.iasri.res.in/ebook/EBADAT/3-Diagnostics%20and%20Remedial%20Measures/5-ROBUST%20REGRESSION1.pdf (accessed December 21, 2011)

[3] Boudjellaba, H., MacGibbon, B. and Sawyer, P., 2001. On exact inference for change in a Poisson

Sequence. Communications in Statistics A: Theory and Methods, 30(3), 407–434.

[4] Chen, J. and Gupta, A. K., 1997. Testing and locating variance change points with application to stock prices. Journal of the American Statistical Association:JASA. 92(438), 739-747.

[5] Chen, J. and Gupta, A. K., 2004. Statistical inference of covariance change points in Gaussian model, Statistics. 38, 17-28.

[6] Fotopoulos, S. B. and Jandhyala, V. K., 2001. Maximum likelihood estimation of a change-point for exponentially distributed random variables. Statistics & Probability Letters. 51, 423-429.

[7] Haccou, P., Meelis, E. and van de Geer, S., 1988. The likelihood ratio test for the change point problem for exponentially distributed random variables. Stochastic Process and Their Applications. 27, 121-139.

[8] Harvey, A. C., 1977. A comparison of preliminary estimators for robust regerssion. Journal of the American Statistical Association. 72(360), 910-913.

[9] Hinich, M, J. and Talwar, P. P., 1975. A simple method for robust regression, Journal of the American Statistical Association. 70(349), 113-119.

[10] Hinkley, D.V., 1970. Inference about the change-point in a sequence of random variables. Biometrika 57(1), 1-17.

[11] Hinkley, D. V. and Hinkley, E. A., 1970. Inference about the chance-point in a sequence of binomial variables, Biometrika. 57(3), 477-488.

[12] Huber, P. J., 1973. Robust regression: Asymptotics, conjectures and monte carlo, Ann. Statist. 1, 799-821.

[13] Jandhyala, V. K. and Fotopoulos, S. B., 1999. Capturing the distributional behaviour of the maximum likelihood estimator of a change point. Biometrika. 86(1), 129–140.

[14] Jandhyala, V. K. and Fotopoulos, S. B., 2001. Rate of convergence of the maximum likelihood estimate of a change-point, Sankhya A, 63(2), 277–285.

[15] Jarrett, R. G., 1979. A note on the intervals between coal-mining disasters. Biometrika. 66(1), 191–193.

[16] Pechenizkiy, M., Bakker, J., Žliobaitė, I., Ivannikov, A. and Karkkainen, T., 2009. Online Mass Flow Prediction in CFB Boilers with Explicit Detection of Sudden Concept Drift. SIGKDD Explorations. 11(2), 109-116.

[17] Ramanayake, A., 2004. Tests for a change point in the shape parameter of gamma random variables. Communications in Statistics A: Theory and Methods. 33, 4, 821-833.

[18] Rousseeuw, P. J. and Leroy, A. M., 1987. Robust Regression and Outlier Detection, Wiley, New York.

[19] Takeuchi, J. I. and Yamanishi, K., 2006. A Unifying framework for detecting outliers and change points from time series, IEEE Transactions on Knowledge and Data Engineering. 18 (4), 482-492.

[20] O'leary, D. P., 1990. Robust regression computation using iteratively reweighted least squares, SIAM J. Matrix Anal. Appl. 11(3),.466-480.

[21] Worsley, K. J., 1986. Confidence region and test for a change-point in a sequence of exponential family random variables. Biometrika. 73(1), 91-104.