AMO- Advanced Modeling and Optimization, Volume 13, Number 3, 2011

Change Point Detection Using Bootstrap Methods

Reza Habibi
Department of Statistics, Central Bank of Iran

Abstract. This paper is concerned with change point detection via bootstrap methods. It is shown that the common marginal distribution of bootstrapped samples is mixture of two distributions before and after the change. We suggest to solve the change point problems as mixture modeling problems. Then, the EM algorithm is given to derive the estimation of parameters of mixture distributions. Some examples show that our algorithm works well. Application of our method in some real data sets is also considered.

Keywords: Bootstrap; Change Point; EM Algorithm; Mixture Distribution

AMS(2000) Subject Classification: 62G20; 62F20.

1 Introduction. During the last four decades, we have witnessed several different methods for detecting change points. Page (1954) studied change point analysis in the context of quality control. Chernoff and Zacks (1964), using a quasi-Baysian approach, modeled the change points. Hinkely (1970) derived the maximum likelihood estimation of change point. Worsley (1988) constructed confidence intervals for change point in the exponential family distributions. An excellent reference in change point problems is Csorgo and Horvath (1997). This topic is a non-regular problem in statistical inference in which the initial distribution of the first k_0 observations, F_{θ_0}, changes at k_0 such that the remaining $n - k_0$ observations come from another distribution say F_{θ_1} where $F_{\theta_0} \neq F_{\theta_1}$. It is known that finding the exact and asymptotic distributions of test statistics and estimators in this field is too difficult. Therefore, the methods of computational statistics are widely used in this research area. Two famous computational methods are Bootstrap and EM algorithm.

The bootstrap is a widely used, computer intensive approach belongs to the class of resampling techniques. It is a sampling with replacement from a given sample. Valid bootstraps are often accessible practical solutions to statistical inference problems. Since Efron (1979) first formally considered the bootstrap technique as another look at the jackknife, there has been a lot of books on bootstrap methods. These have addressed regular and non-regular problems and have allowed the observations to be either independent or dependent, see Efron and Tibshirani (1993), Davison and Hinkley (1997) and Lahiri (2003)

[1]Reza habibi is staff of Central Bank of Iran. He has a PhD in Statistics form Shiraz U.

***AMO - Advanced Modeling and Optimization. ISSN: 1841-4311**

among the others. The EM algorithm is a useful method to derive the maximum likelihood of a parameter numerically, developed by Dempster *et al.* (1977). It is used in non-regular distributions like the mixture distributions.

In this paper, in a sequence with a change point, we derive the bootstrap samples. It is seen that the common marginal distribution of bootstraped samples is mixture of two distributions, i.e., before and after change point with mixing proportion k_0/n. Therefore, we use the EM algorithm to estimate the parameters of mixture distribution, we mean θ_0, θ_1 and k_0. In fact, we solve the change point problems using statistical tools applied to the mixture distribution topic. The rest of this paper is organized as follow. Section 2 is involved by the application of bootstrap methods in change point analysis. In section 3, we consider some examples. The change point detection in some real data sets is considered in section 4. Throughout of this paper, we assume that $X_1, ..., X_{k_0}, X_{k_0+1}, ..., X_n$ is a sequence of independent random variables. The k_0 is unknown change point. Variables $X_1, ..., X_{k_0}$ are independent and identically distributed (*iid*) form F_{θ_0} and the remaining X_{k_0+1}, ..., X_n come from F_{θ_1}.

2 Bootstrap in change point. The bootstrap methods has many applications in change point analysis. In the case of *iid* sequence (without change point) of random variables $X_1, ..., X_n$, under he usual bootstrap approach, we derive a sample $X_1^*, ..., X_n^*$ with replacement and with equal probabilities $1/n$ assigned to X_is. This equivalent to deriving samples $X_1^*, ..., X_n^*$ from F_n, the empirical distribution of the original observations. For a sequence of random variables with a change point, the following three steps are considered.

1. Using a suitable estimation technique like cusum, cusumsq,... estimate k_0, θ_0 and θ_1 respectively (see Csorgo and Horvath, 1997).

2. Apply twice bootstrap approach for sequences $X_1, ..., X_{\widehat{k}_0}$ and $X_{\widehat{k}_0+1}$, ..., X_n, independently. To this end, it is enough to generate subsamples from $F_{\widehat{\theta}_0}$ and $F_{\widehat{\theta}_1}$, respectively. When the functional form is known, this is a parametric bootstrap. For unknown functional forms of F_{θ_0} and F_{θ_1}, they are substituted by their empirical distribution functions. This is a nonparametric bootstrap approach.

3. Apply the above mentioned (in step 1) estimation methods to bootstrapped samples $X_1^*, ..., X_{\widehat{k}_0}^*$ and $X_{\widehat{k}_0+1}^*, ..., X_n^*$ to derive $\widehat{k}_0^*$, and then use $X_1^*, ..., X_{\widehat{k}_0^*}^*, X_{\widehat{k}_0^*+1}^*, ..., X_n^*$ to receive to $\widehat{\theta}_0^*$ and $\widehat{\theta}_1^*$, respectively.

Here, we suggest using this simple bootstrap (don't consider to steps 1-3) for a sequence of independent observations with a change point an unknown time

point k_0, that is for $\mathbf{X} = (X_1, ..., X_{k_0}, X_{k_0+1}, ..., X_n)$. It is seen

$$\begin{aligned} F_{X_i^*|\mathbf{X}}(x) &= E\{\mathbf{1}(X_i^* \leq x)|\mathbf{X}\} = F_n(x) \\ &= (1/n)\sum_{i=1}^{n} \mathbf{1}(X_i \leq x), \end{aligned}$$

where $\mathbf{1}(a \leq b)$ is one if $a \leq b$ and zero otherwise. By taking expectation from both sides of above equation, we find that

$$\begin{aligned} F_{X_i^*}(x) &= (1/n)\sum_{i=1}^{n} P(X_i \leq x) \\ &= \frac{k_0}{n} F_{\theta_0}(x) + (1 - \frac{k_0}{n}) F_{\theta_1}(x), \end{aligned}$$

that is, the marginal distribution of bootstraped independent samples $X_1^*, ..., X_n^*$ is mixture of two distributions (that is distributions before and after the change point) with mixing proportion $k_0/n = 1/n, ..., (n-1)/n$.

It is generally believed that working with mixture distributions is much easier than to work with change point problems. For example, one can compare the corresponding asymptotic distributions. The asymptotic distributions in change point fields are functional of stochastic processes like Brownian motion whereas these distributions are often normal in mixture distribution. This is why, we change our shift point problem to mixture distribution to make inference about θ_0, θ_1 and k_0. A standard approach for estimating the parameter of mixture distributions is EM algorithm (see Dempster *et al.*, 1977). Unfortunately, the predefined commands of statistical packages like SAS or SPLUS for EM algorithm aren't suitable here since the mixing proportion in our problem takes only finite values $1/n, ..., (n-1)/n$. To overcome this difficulty, we advise to run the EM algorithm $n-1$ times for each cases $k_0 = 1, ..., n-1$. In each step, we compute $\widehat{L}(k_0)$

$$\widehat{L}(k_0) = \prod_{i=1}^{n} (\pi_0 f_{\widehat{\theta}_0(k_0)}(x_i) + (1 - \pi_0) f_{\widehat{\theta}_1(k_0)}(x_i)),$$

with $\pi_0 = \frac{k_0}{n}$ at which $\widehat{\theta}_0(k_0)$, f_{θ_0} and $\widehat{\theta}_1(k_0)$, f_{θ_1} are the estimations of parameters and densities before and after fixed change point k_0, respectively. Then $\widehat{k}_0$, $\widehat{\theta}_0$ and $\widehat{\theta}_1$ are the maximizer of $\widehat{L}(k_0)$, $\widehat{\theta}_0(k_0)$ and $\widehat{\theta}_1(k_0)$, respectively. To maximize $L(k_0)$ for each fixed k_0 with respect to $\theta_0(k_0)$ and $\theta_1(k_0)$ we apply the EM algorithm. If we repeat the bootstrap procedure R (sufficiently large) times, we can obtain the sampling properties of $\widehat{k}_0$, $\widehat{\theta}_0$ and $\widehat{\theta}_1$ even for small sample sizes n. Hypotheses $k_0 = n$ or $\theta_0 - \theta_1 = 0$ stands for null hypothesis of no change point. The sampling distributions of $\widehat{k}_0$, $\widehat{\theta}_0 - \widehat{\theta}_1$ are useful tools to

test the null hypothesis. Note that to test $H_0 : k_0 = n$, we should let k_0 moves among 1 to n.

We insist that the marginal distribution of X_i^* is mixture and X_i^* itself is obtained from conditional distribution given sample $\mathbf{X} = (X_1, ..., X_{k_0}, X_{k_0+1}, ..., X_n)$. To generate samples $Y_1, ..., Y_n$ distributed as $\frac{k_0}{n}F_{\theta_0} + (1 - \frac{k_0}{n})F_{\theta_1}$, in practice, we propose the following scheme: generate n *iid* indices $I_1, ..., I_n$ among $\{1, 2, ..., n\}$ independent of $\mathbf{X}$ and let $Y_i = X_{I_i}$, $i = 1, 2, ..., n$. Then, note that

$$\begin{aligned} F_{Y_i}(y) &= P(Y_i \leq y) = \sum_{j=1}^{n} P(X_{I_i} \leq y | I_i = j)/n \\ &= \sum_{j=1}^{n} P(X_j \leq y)/n \\ &= \frac{k_0}{n}F_{\theta_0}(y) + (1 - \frac{k_0}{n})F_{\theta_1}(y). \end{aligned}$$

We summarize our approach in the following steps. Again, consider $\mathbf{X} = (X_1, ..., X_{k_0}, X_{k_0+1}, ..., X_n)$.

1. Generate an *iid* sequence (with replacement) $I_1, ..., I_n$ among $\{1, 2, ..., n\}$ independent of $\mathbf{X}$ and let $Y_i = X_{I_i}$, $i = 1, 2, ..., n$.

2. For each $k_0 = 1, ..., n-1$, run $n-1$ EM algorithms using a realization of $Y_1, ..., Y_n$ (taken from 1) to obtain $\widehat{L}(k_0)$, $\widehat{\theta}_0(k_0)$ and $\widehat{\theta}_1(k_0)$. The argmax of these quantities as functions of k_0 are $\widehat{k}_0$, $\widehat{\theta}_0$ and $\widehat{\theta}_1$.

3. Repeat steps 1-3 for R times to obtain $\widehat{k}_0^{(r)}$, $\widehat{\theta}_0^{(r)}$ and $\widehat{\theta}_1^{(r)}$, $r = 1, 2, ...,$R. In this way, the sampling properties of $\widehat{k}_0$, $\widehat{\theta}_0$ and $\widehat{\theta}_1$ are derived.

3 Examples. Here, we propose some examples. In a sequence of size 100 of independent observations, there is a change point in k_0. The change point estimate, $\widehat{k}_0$, and the estimations of parameters before and after change point, $\widehat{\theta}_0$ and $\widehat{\theta}_1$, are given. The results are given in the following Table. Notations N, Exp, t, C stand for normal, exponential, t-student, and Cauchy distributions, respectively. The following Table (following the rows) shows that our method works well in these cases (1) when the mean increases after the change point, (2) mean decreases, (3) the variance changes, (4) variance and mean change at the same time, (5) the parameter of distribution increases (in exponential distribution), (6) the normal distribution shifts to Cauchy distribution with different parameters, (7) changes in degrees of freedom in t distribution, (8) the parameter of distribution decreases (in exponential distribution) and change point is close to the end of sequence, (9) the support of observations changes

from $(-\infty, \infty)$ to $(0, \infty)$, and (10) means changes and change point is too close to the end of sequence.

Table 1: Simulation Results

dist(before)	dist(after)	k_0	$\widehat{k}_0$	$\widehat{\theta}_0$	$\widehat{\theta}_1$
$N(\theta_0 = 0, 1)$	$N(\theta_1 = 2, 1)$	25	21	0.06	1.85
$N(\theta_0 = 1, 1)$	$N(\theta_1 = -1, 1)$	35	32	1.77	-0.95
$N(0, \theta_0 = 1)$	$N(0, \theta_1 = 2)$	40	39	1.25	2.03
$N(0, 1)$, θ_0= (0,1)	$N(2, 3)$, θ_0= (2,3)	45	41	(0.05,0.98)	(1.77,3.12)
$Exp(\theta_0 = 1)$	$Exp(\theta_1 = 3)$	50	49	1.12	2.77
$N(\theta_0 = 0, 1)$	$C(\theta_1 = 1, 1)$	50	44	0.04	1.45
$t_{\theta_0=2}$	$t_{\theta_1=3}$	55	50	2.4	3.3
$Exp(\theta_0 = 1)$	$Exp(\theta_1 = 1.5)$	75	77	1.05	1.35
$N(\theta_0 = 0, 1)$	$Exp(\theta_1 = 1)$	85	88	0.05	1.05
$N(\theta_0 = 0, 1)$	$N(\theta_1 = -2, 1)$	93	90	0.15	-1.75

4 Real data sets. Here, we study the performance of our method in some before examined data sets having a change point.

4.1 Stock market data. The data set (taken from Hsu (1979)) is the weekly log price relative (X_i) of the Dow Jones Industrial Average for the period July 1, 1979 to August 2, 1974. Hsu (1979) has shown that X_i's are independent and zero mean and normally distributed. However, there can be doubts about the constancy of variances of the sequence. It is seen that the sequences if more variable in the later periods than the earlier. To check this possibility, we applied our method and found that $\widehat{k}_n = 89$ is the point at which the variances differ.

4.2 Aircraft arrival times. There are 212 inter aircraft arrival times within noon through 8 p.m. on April 30 1969. The observations are independent and have exponential distribution (see Hsu (1979)). He has shown that the data have no change point. To examine this, we applied our method to data set. It is seen that the change point estimation is 211. Therefore, we found that the arrivals have a coomon rate but this rate is changes for the last observation.

4.3 Heart transplant data. This data set was taken from Kalbfleisch and Prentice (1980). The average survival time for 35 known age groups are considered. They have proved that the data are distributed exponentially. We understood that there is a change point in $k_0 = 11$ and the parameters before and after the change point are $\theta_0 = 202$ and $\theta_1 = 368.8$.

Remark 1. We applied our method in Simulated data (Page,1954), Nile River Data (Cobb,1978) and Pettitt's data (Pettitt,1979). We saw that our

method works well in all above mentioned data sets. The results are npt given here. Interested readers can refer to Habibi (2010).

References

[1] Chernoff, H., Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Statist.* **35**, 999–1018.

[2] Cobb, G.W. (1978). The problem of the Nile: conditional solution to a change-point problem. *Biometrika* **65**, 243-251.

[3] Csorgo, M., Horvath, L. (1997). *Limit theorems in change point analysis.* Wiley, New York.

[4] Davison, A. C, Hinkley, D.V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

[5] Dempster, A.P, Laird, N.M and Rubin, D.B. (1977). Maximum Likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society*. Series B, **39**, 1-38.

[6] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

[7] Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

[8] Habibi, R. (2010). Change point detection using bootstrap methods. *Tech Report*. Department of Statistics, Central Bank of Iran.

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

[10] Hsu, D.A. (1979). Detecting shifts of parameter in gamma sequences, with applications to stock price and air traffic flow analysis. *J. Amer. Statist. Assoc.* **74**, 31-40.

[11] Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[12] Lahiri, S. N. (2003). *Resampling methods for dependent data* . Springer-Verlag , New York.

[13] Page ES (1954). Continuous inspection schemes. *Biometrika* **41**, 100–115.

[14] Pettitt, A.N. (1979). A non-parametric approach to the change-point problem. *Applied Statistics* **28**, 126-135.

[15] Worsley KJ (1986). Confidence regions and test for a change point in a sequence of exponential family random variables. *Biometrika* **73**, 91- 104.