Bangladesh J. Sci. Res. 25(1): 87-92, 2012 (June) **Short communication**

DETECTING JUMPS IN CORRELATED SERIES

M. Shahidul Islam*

Department of Statistics, Shahjalal University of Science & Technology,
Sylhet, Bangladesh

This paper addresses the issue of testing mean for a correlated series, preferably in genomic data. In the proposed method, the order of the autoregressive model in a given segment, referring to a series, is detected and hence a parametric bootstrap procedure is applied. However, if the segment of interest is not preselected, we propose using Maximum Overlap Discrete Wavelet Transform (MODWT) for detecting the change points of interest. The performance of the method is evaluated and compared with other conventional method using simulation procedure. The applicability of the method is demonstrated through a real genomic data.

In the study of time series, it is quite often necessary to detect the change point. For example, array CGH (Comparative Genomic Hybridization) is a molecular-cytogenetic method for genomewide screening to detect chromosome gains or losses in the DNA content (Pinkel and Albertson 2005, Willenbrock and Fridlyand 2005). One of the recent methods using MODWT (Percival and Walden 2000) can be used to detect change points in a series (Islam 2008). In order to find the possible points of interest, we can use R package Wave CD which is available at CRAN (Islam 2010). The method is incomplete in the sense that it checks for jumps compared to the previous level. However, in the aforementioned case, our goal is to test whether the mean of any section is significantly different from zero.

As the series have correlation among the successive observations, the regular *t* test is not applicable. Kim (1996) used likelihood ratio test for testing mean for such a correlated series. The method of detecting the order of an ARMA process has been described by McLeod and Zhang (2007). The software is also freely available online. We adopt this package, namely FitARMA, in finding the structure of the series. Once this is done, simulation technique can be incorporated to generate a series with the detected structure. This method is formally called parametric bootstrapping (Davison and Hinkley 1997; Efron and Tibshirani 1993). We generate the series very large number of times from a mean-zero model and find the number of means more extreme than the observed one. This simply refers to the p value for the test of mean different from zero. It is demonstrated through simulation that the power of such test is much higher than the method with corrected standard deviation. Finally, we apply this method to real

*Corresponding author. E-mail: shahed_sta@surt.edu

microarray data to call a chromosomal region as loss or gain region.

Suppose, we have z_t, $t = 1, 2, ..., n$ as the observations along a specific chromosome arm. The observations in ith region and tth position can be expressed as

$$z_{ti} = \mu_i + e_t\ ,\ i = 1, 2, ..., k \text{ and } t = 1, 2, ..., n$$

The error term e_t follows AR(p) process. That is,

$$e_t\ = \varphi_1 e_{t-1} + \varphi_2 e_{t-2} + ... + \varphi_p e_{t-p} + a_t \qquad (1)$$

where $\varphi_1, \varphi_2, ..., \varphi_p$ are autoregressive parameters and $e_t \sim N(0, \sigma_a^2)$.

For simplicity, let us consider that we have only one region and we would like to test whether the region mean is significantly different from zero. Intuitively, we can think about correcting the standard deviation in denominator of the t statistic using the modified formula in a correlated series with autoregressive moving average process of order p and q. A t-test procedure that considers corrected variance of $\bar{z}$ in an ARMA (p, q) process would seem to work for such case. Unfortunately, this intuitive method fails to maintain a standard power of the test, which is described in next section.

To overcome lack of power of the test in such phenomenon, we can resort to parametric bootstrapping procedure. This simple method can be outlined in the following few steps:

Step 1: If the regions are defined, find $e_{ti} = y_{ti} - \hat{y}_i$. However, if the regions are not defined, find the breaks points using MODWT procedure or some other method and hence find $e_{ti} = y_{ti} - \hat{y}_i$.

Step 2: Select the AR order p from the series obtained in step 1.

Step 3: Estimate the parameters and innovation variance from model selected in step 2.

Step 4: Simulate a mean-zero stationary Gaussian AR(p) time series, say e^*, with parameters $\hat{\phi}$ and innovation variance $\hat{\sigma}$ found in step 3. For null model $\mu = 0$, and so $y = e$. Do the simulation procedure large number of times, say $B = 10^4$ times.

Step 5: Find the means for each simulated series in all regions, $\bar{y}_{j1}{}^*$, $\bar{y}_{j2}{}^*$, ..., $\bar{y}_{jk}{}^*$ where $\bar{y}_{ji}{}^*$ denotes the mean for region i in jth bootstrap sample. The p value for region i is defined as, $p_i = \#\{\bar{y}_{ji}{}^* \geq \bar{y}_i\}/B$.

To evaluate the power of this test procedure, a short simulation study with an AR(1) process was done. The results are presented in the following table, which has two parts corresponding to $\mu = 0$ and $\mu = 0.5$. It is worth mentioning that $\mu = 0$ refers to probability of type-I error and $\mu = 0.5$ refers to power against one single point 0.5. The first part of each column of the table reveals that the conventional method does not perform very well even for small φ values. Hence with the increase of magnitude of φ, the method becomes incapable of handling such situation regardless of the series length.

The simulation study, presented in the second part of each column in the table, suggests that the bootstrapping method works well for testing mean in large series. The *False Positive Rate* (FPR), described in Benjamini and Hochberg (1995), is still high for large φ and short series. However, series length refers to the length of a particular chromosome, which in real CGH data will be moderate to large.

Table 1. Power of the test $\mu = 0$ in an AR(p) setting with series length 50 and 100. The first part of the column for $\mu = 0$ and $\mu = 0.5$ represents the test with conventional t-test with corrected standard deviation. The second part represents results from bootstrap approach. Here, the column for $\mu = 0$ represents type-I error and the test is done at 0.05 level of significance. For all cases, we consider standard deviation for error term to be 0.2.

Power comparison; $n = 50$, $\sigma_a = 0.2$			Power comparison; $n = 100$, $\sigma_a = 0.2$		
φ	$\mu = 0$	$\mu = 0.5$	φ	$\mu = 0$	$\mu = 0.5$
0.0	0.056|0.064	0.942|1.00	0.0	0.055|0.054	0.999|1.00
0.1	0.084|0.064	0.907|1.00	0.1	0.072|0.056	0.995|1.00
0.3	0.118|0.066	0.747|1.00	0.3	0.078|0.064	0.941|1.00
0.5	0.108|0.080	0.519|1.00	0.5	0.079|0.062	0.722|1.00
0.7	0.137|0.118	0.310|0.99	0.7	0.097|0.072	0.396|1.00
0.9	0.236|0.244	0.286|0.75	0.9	0.166|0.134	0.201|0.83

The implementation of the method was performed using a CGH array where 2400 BAC (Bacterial Artificial Chromosome) clones were measured each with three replicates (Snijders *et al.* 2001). Measurements for log base 2 intensity ratio are provided and these are considered as a time series sequence.

Average relative DNA copy number sequences of the three replicates along the genome is shown in the following Figure. Different change points were detected using the R package WaveCD. Then the means were tested using the proposed method. If any region has a real jump significantly away from zero, this is colored as red or green depending on gain region or loss region respectively. These regions are called abnormal regions along the chromosomes but the other regions which are colored as yellow or blue are not detected as abnormal regions. As we can see, the measures are mostly along the zero line, which indicates that the test sample has the same DNA copy numbers as that of reference sample. Overall, the presence of abnormal regions can be observed in several chromosomes, namely 1, 5, 7, 8, 9, 11, 14, 17, 20 and 23.

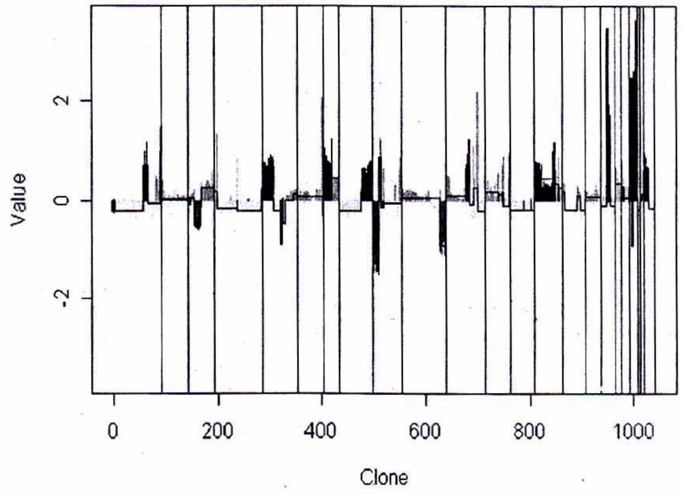

Figure 1. Application of the proposed bootstrap method to CGH data set from Snijders *et al.*· (2001). Possible break points are detected using R package WaveCD. If the mean for any segment is significantly different from zero, then this is marked as red or green depending on whether it is gain or loss region respectively. The normal regions are colored as yellow and blue. Different chromosomes are detected to have true gain and loss regions.

If there is only one region presents in the study, the decision about the test can be done using this obtained p value. However, in an array CGH data there will be several regions of interest and so the overall decision depends on multiple test method. Having obtained the p values for all regions using the aforementioned bootstrap procedure, we can calculate the multiple test values using some standard method. Benjamini and Hochberg (1995) proposed a method for multiple testing using *False Discovery Rate* (FDR). Another more recent approach, called q-value, was proposed by Storey (2002). However. unlike the number of genes, the number of jump points or the number of regions will not be even hundreds. So it would be expected that these methods would produce similar results in this simulation.

We have presented a simple but effective method for calling true gain or loss region in a CGH array. Although the simulation was done for simple AR(1) process, this can be extended to higher order and also pretty many combinations of φ, σ and series length. The results reveal that this parametric bootstrap approach has much higher power than the conventional method. Although the performance gets better with the increase of the size of the series, it is still applicable for short series length. The method is flexible for jump detection in any time series. To implement this proposed method, the use of multiple test method is recommended in case of many regions detected using the R package Wave CD along a chromosome. The R codes for implementing the method would be available on request.

References

Benjamini Y. and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statist. Soc. B,* **57:** 289-300.

Davison, A.C. and Hinkley, D.V. 1997. *Bootstrap Methods and Their Application*, Cambridge University Press.

Efron, B. and Tibshirani, R. 1993. *An Introduction to the Bootstrap*, Chapman & Hall.

Islam, M. 2008. Periodicity, Change Detection and Prediction in Microarrays. Ph.D. Thesis, University of Western Ontario.

Islam, M. S. 2010. WaveCD: Wavelet change point detection for array CGH data [url= http://CRAN.R-project.org/package=WaveCD].

Kim, H-J. 1996. Change-point detection for correlated observations. *Statistica Sinica* **6**: 275-287.

McLeod, A.I. and Zhang, Y. 2007. Faster ARMA maximum likelihood estimation, Computational Statistics & Data Analysis 52(4), URL http://dx.doi.org/ 10.1016/j.csda.2007.07.020

Percival, D. B. and Walden, A. T. 2000. *Wavelet Methods for Time Series Analysis.* Cambridge University Press.

Pinkel, D. and Albertson, D. G. 2005. Array comparative hybridization and its applications in cancer. *Nature Genet-ics*, **37**: S11-S17.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown N., Conroy, J., Hamilton, G., Hindle, A. K, Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D. and Albertson, D. 2001. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29:** 263 - 264.

Storey, J. D. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**: 479-498.

Willenbrock, H. and Fridlyand, J. 2005. A comparison study: applying segmentation to array cgh data for down-stream analyses, *Bioinformatics* **21**(22): 4084-4091.

(Manuscript received on 26 April 2012; revised on 28 July 2012)