Far East Journal of Theoretical Statistics

Published Online: August 2015
http://dx.doi.org/10.17654/FJTSMay2015_181_190
Volume 50, Number 3, 2015, Pages 181-190 **ISSN: 0972-0863**

SPECTRAL ANALYSIS OF PERIODICALLY BEHAVING BIOLOGICAL TIME SERIES

B. Onoghojobi, N. P. Olewuezi and I. M. Okeke

Department of Statistics
Federal University of Technology Owerri
Nigeria

Abstract

Periodic behavior has been a successful feature in cellular processes and biological processes that are highly contaminated with noise problem. In this paper, we introduce the use of a new test procedure for finding periodic sequence in multiple series or microarray data, with respect to hypothesis testing method based on the Fisher *g*-statistic. Using the basic idea underlying a robust testing procedure under the Gaussian noise assumption, a computational efficiency means of detecting periodicity in biological time series data was established.

1. Introduction

Periodic phenomena are widespread in biology. The problem of finding periodicity in biological time series can be viewed as a multiple hypothesis testing of the spectral content of a given time series. The exact noise characteristics are unknown in many bioinformatics applications.

Received: January 8, 2015; Revised: March 21, 2015; Accepted: April 23, 2015
2010 Mathematics Subject Classification: Primary 62P10.
Keywords and phrases: periodic behavior, cellular processes, periodic sequence, Fisher *g*-statistic.
Communicated by K. K. Azad

Furthermore, the observed time series can exhibit other non-idealities, such as outliers, short length and distortion from the original wave form. Hence, the computational methods should preferably be robust against such anomalies in the data.

This approach can be parametric or non-parametric. The parametric approaches assume that the underlying stationary stochastic process has a certain structure which can be described using a small number of parameters. The major task in this method is to estimate the parameters of the model that describes the stochastic process.

The non-parametric approaches explicitly estimate the covariance or the spectrum of the process without assuming that the process has any particular structure.

DNA microarray experiments are usually classified based on the type of array that is used in the experiment (cDNA and oligonucleotide arrays) or according to the organism that is profiled. In static expression experiments, a snapshot of the expression of genes in different samples is measured, while in time series expression experiments, a temporal process is measured. Another important difference between these two types of data is that while static data from a sample population (e.g., ovarian cancer patients) are assumed to be independent identically distributed; time series data exhibit a strong autocorrelation between successive points.

Much of the early work on analyzing time series expression experiments used methods developed originally for static data [1, 3]. More recently, several new algorithms specifically targeting time series expression data were presented in literature.

The problems of processing gene expression and other molecular biological time series data include short time series length, the presence of noise of unknown distribution, outliers and other non-linearities involved in measurement technologies themselves. In [4], a robust rank-based modification of Fisher's *g* test for finding hidden periodicities in time series data was established. The method performs well both under the Gaussian

noise assumption and when outliers and other non-linearities are present. The method, however, requires intensive numerical computation when it comes to evaluating the significant values.

A major difference between the methods has been proposed [2]. Wichert's method is capable of detecting unknown frequencies whereas other methods are designed for detecting fixed frequencies. From a computational point of view, the problem of finding unknown frequencies is even more demanding since no prior knowledge of the frequency to be detected is available.

The main contribution of this paper is to establish methods of detecting periodicity in biological processes and comparing their relative efficiency.

In Section 2, we provide some background on the methods employed in the detection of periodicity and also in the analysis of biological processes. In Section 3, we discuss on the analysis and result of our experiment on all the biological time series data.

2. Methodology

2.1. Method of analysis

In order to be consistent with the previously published methods, we use similar notation as in [2] and also consider the same model for the periodic time series

$$Y_n = \beta \cos(\omega n + \phi) + \varepsilon_n, \tag{1}$$

where $\beta > 0$, $\omega \in (0, \pi)$, $n = 1, ..., N$, $\phi \in (-\pi, \pi]$, and ε_n is an i.i.d. noise sequence. To test for the periodicity, define the null hypothesis as $H_o : \beta = 0$, i.e., time series consists of the noise sequence alone, $y_n = \varepsilon_n$. We review a method for detecting unknown frequencies and later introduce a modification which can be applied to the detection of known frequencies. We first review the Fisher's test for the detection of periodic transcripts as introduced in [2].

The method proposed by [2] is based on the periodogram spectral estimator defined as

$$I(\omega) = \frac{1}{N}\left|\sum_{n=1}^{N} y_n e^{-i\omega n}\right|^2, \quad \omega \in [0, \pi], \tag{2}$$

where N is the time series length. The periodogram is further evaluated at (harmonic) normalised frequencies

$$\omega_l = \frac{2\pi l}{N}, \quad l = 0, 1, ..., a, \tag{3}$$

where $a = [(N-1)/2]$ and $[x]$ denotes the integer part of x. To test for the periodicity formally, some kind of a test statistic must be chosen, hence the use of the Fisher g-statistic. The so-called g-statistic for one time series is given by

$$g = \frac{\max_{1\leq 1\leq a} I(\omega_l)}{\sum_{i=1}^{a} I(\omega_l)}.$$

In plain words, the g-statistic is the maximum periodogram ordinate divided by the sum of all periodogram ordinates for $I = 1, ..., a$. Large value of g indicates a strong periodic component and leads to the rejection of the null hypothesis.

Wichert et al. [2] resorted to a result by Fisher, under the Gaussian noise assumption that gives the exact distribution of the g-statistic with respect to its null hypothesis.

To correct the p-values for multiple testing, use the method of Benjamini and Hochberg, which controls the false discovery rate (FDR). The FDR method controls the expected proportion of false positives (type I errors) at a given rate q. The threshold of the FDR depends on the evaluated p-values.

However, the same methods, such as simulation and permutation-based significance values, can also be applied to the modified g-statistic. In experimental results section, we apply both the standard and the modified g-statistics to real microarray data.

2.2. Simulation and permutation

The simulation-based method is simple. Given the model as in equation (1) together with some distributional assumptions for ε_n, generate a set of P random time series under the null hypothesis. Use the obtained g-values to compute an estimate of the distribution of the g-statistic under the null hypothesis. The distribution can be estimated, e.g., using kernel density estimation methods. The testing can then be performed as explained above except that the significance values are computed/integrated relative to the estimated distribution.

Note that the null distribution must be estimated for each time series length separately but, due to the distribution-free property, the null distribution is independent of the noise characteristics under the i.i.d. assumption.

A more flexible way of obtaining p-values is to use permutation tests. Although they are a relatively old concept, permutation tests have only recently become interesting in practice because of the intensity of needed computing power. The idea is simple:

Choose a test statistic.

Evaluate the test statistic on the original data.

Randomly permute the data and evaluate the test statistic on every permutation.

Estimate the distribution of the test statistic with the help of the sample generated.

Use the estimated distribution to get a p-value for the original test statistic computed.

A sequence of random variables $\{X_n\}$, $n = 1, 2, ..., N$ is exchangeable, if the joint distribution of $X_{\pi 1}, X_{\pi 2}, ..., X_{\pi N}$ is the same as that of the original sequence $X_1, X_2, ..., X_N$ for all permutations π. Under the null

hypothesis, the elements of the time series y_n are independent and identically distributed. And therefore exchangeable, and hence the permutation test can be applied.

Alternatively, as the application of a random permutation destroys any periodic structure that is present in the original sequence, permutation tests can be used to assess how highly structured the given time point values are in the light of the chosen test statistic versus other permutations of the given sample. As the concept of permutation tests is non-parametric, they can be applied without knowing the exact distribution of the data at hand.

Instead of performing the entire *N*! permutations for each time series, we have chosen to permute each of the original time series for $P = 5000$ times. As our simulations show, this seems to be quite an adequate number of iterations.

Let us first examine the power of the test, i.e., one minus the probability of the type II error (false negative). The power of the test is estimated for the three different test cases as well as for different time series lengths and for different noise parameters using 10000 Monte Carlo runs. The significance level is set to $\alpha = 0.05$. In all the three cases, the case-specific noise assumptions are used for both the null hypothesis $(\beta = 0)$ and the alternative hypothesis $(\beta > 0)$. In this simulation, we use the signal model to represent a periodic signal (i.e., the alternative hypothesis). In the right column of Table 3, the length of the time series is set to 40 and the power is shown as the function of varying noise parameters. Table 3 clearly shows that the power of the proposed robust hypothesis testing method is remarkably better than that of the Fisher's test, especially in the case of outliers and non-linear distortion. More interestingly, however, the proposed method is also more powerful in the case of standard Gaussian noise.

Next, we consider another simulation. In the same way as in [2], two thousand time series of length $N = 10, 20, 40, 45, 50$ and 100 were generated to test the periodicity detection. One thousand and nine hundred of

the time series were plain noise and one hundred time series were generated. We again consider the three aforementioned noise models. As explained in the computational methods subsection, we evaluated the *g*-statistic and *p*-value for each time series and then used the FDR rule to determine which of the time series was considered to be cyclic for a certain FDR level. The FDR level, at which the expected rate of false positives is controlled, was chosen similarly as in [2], i.e., $q = 0.15, 0.10, 0.05, 0.01$ and 0.005. For each *N* and *q*, the simulation was run for 99 times for the simulation-based cases and 9 times for the permutation-based cases. Median statistics are reported for the number of found periodic components, the number of correctly identified periodic components (shown in parenthesis) and the number of truly periodic time series among the top 100 ranked sequences (Z).

If we take a look at the results in Tables 1 to 3, we can draw some immediate conclusions:

There are no significant differences between the two methods in terms of the number of detected genes or in terms of the number of correctly detected genes. However, the numbers of truly periodic genes among the top 100 ranked sequences (*Z*-scores) show somewhat favorable performance for the robust method, especially for the short time series $N = 20$ and $N = 40$. Indeed, this observation agrees with previous findings [3] where the robust method was found to have a good performance as a spectrum estimator for short time series. By comparing Tables 1, 2 and 3, it is obvious that the permutation tests do not provide any significant performance gain over the traditional approach where the significance values are computed using the simulation-based method, respectively. In both cases, the *Z*-scores are about the same, as expected. The only notable difference is seen in the number of found periodic genes for short time series $(\text{e.g., } N = 40, 45, 50)$ and small FDR levels $(q = 0.005, 0.01, 0.05)$, where the numbers are slightly higher when permutation tests are used. This suggests that the permutation-based method finds a bit smaller *p*-values than the simulation-based method.

3. Data Analysis and Result

For each time series experiment (13 in total), the proposed robust methods for detecting genes having both fixed and unknown frequency components was applied. For the fixed frequency, we use the one that corresponds to the length of the cell cycle. Following the idea presented by Wichert et al., a simple method for estimating the cell cycle length/frequency is to compute the average robust spectral estimate. For each time series, we present the number of statistically significant genes that are found to be periodically behaving at a specific level of the FDR $(q = 0.05)$. For the Spellman method, the sampling time was not equidistant in the beginning and at the end of the data set.

Any monotone distortion preserves the ordering of the samples. Therefore, the rank-based method is completely insensitive to any monotone distortions. Consequently, the results for the third test case are identical to those presented in Tables 1 and 2. The results for the periodogram method are shown in Table 3.

Table 1. Number of inferred periodic time series: The number of inferred periodic time series using the robust method and standard Gaussian noise in the data. *P*-values were obtained by simulating the distribution of the *g*-statistic using 10000 time series composed of Gaussian noise

$q\backslash N$	10	20	40	45	50	100
0.15	0	2(1)	107(90)	109(96)	117(99)	115(100)
0.10	0	1(1)	96(87)	103(94)	110(98)	109(100)
0.05	0	1(0)	83(79)	95(89)	101(96)	105(100)
0.01	0	1(0)	59(59)	80(79)	90(89)	101(100)
0.005	0	0(0)	32(32)	62(61)	64(64)	100(100)
Z	12	49	89	93	95	100

Table 2. Number of inferred periodic time series: The number of inferred periodic time series using the robust method and standard Gaussian noise in the data. *P*-values were obtained by using permutation tests

*q**N*	10	20	40	45	50	100
0.15	0	4(3)	108(92)	113(96)	111(98)	119(100)
0.10	0	1(1)	99(90)	106(94)	106(97)	112(100)
0.05	0	1(0)	88(84)	97(89)	101(95)	106(100)
0.01	0	0	65(64)	80(78)	86(86)	101(100)
0.005	0	0	46(46)	61(61)	71(71)	100(100)
Z	15	48	91	92	95	100

Table 3. Number of inferred periodic time series: The number of inferred periodic time series using the periodogram method and standard Gaussian noise and cubic distortion in the data. *P*-values were obtained by using permutation tests

*q**N*	10	20	40	45	50	100
0.15	0	0	49(44)	79(64)	89(74)	107(93)
0.10	0	0	39(36)	71(62)	80(69)	98(90)
0.05	0	0	25(24)	52(49)	64(59)	90(88)
0.01	0	0	8(8)	28(28)	44(43)	82(82)
0.005	0	0	8(8)	19(19)	37(36)	67(67)
Z	7	15	68	71	79	91

4. Implication of Results

Although the theoretical null distribution was derived by Fisher for Gaussian noise, we found that deviation from it can be significant when signal length is short. Moreover, when the signal does not cover an integer number of periods, significant drop in the statistical power of the test was observed. In this case, a much longer signal is needed for the test to return

reliable result. We found that Fisher test is relatively robust to noise. We also investigate how the FDR multiple testing correction strategy affects the number of detected periodic signals. Although the Fisher test may be unreliable for short signal, the Fisher *g*-statistic has been observed to provide a useful ranking of periodic signals.

All these findings have important implications for periodic gene expression profiles detection as these profiles are often noisy, of very short length, and often with unknown periodicity. In high likelihood, the number of periodic gene expression profiles can be severely underestimated for short length signal as is found with many of the publicly available gene expression datasets. The presented method yields a robust way of finding periodicity in short time series data. As illustrated in simulations Subsection 2.2, the proposed robust detection method is remarkably insensitive to different kinds of non-idealities in the data, such as heavy contamination of outliers, missing values, short time series, non-linear distortions, and is completely insensitive to any monotone non-linear distortions. The results also show that the proposed method has clearly better performance than the Fisher's test, even in the case of the standard Gaussian noise. Furthermore, the results on real data demonstrate that the proposed method performs well on real data and that the results are biologically meaningful.

References

[1] M. Schena, D. Shalon, R. Davis and P. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995), 467-470.

[2] S. Wichert, K. Fokianos and K. Strimmer, Identifying periodically expressed transcripts in microarray time series data, Bioinformatics 20 (2004), 5-20.

[3] L. Tatum and C. Hurvich, High breakdown methods of time series analysis, Journal of the Royal Statistical Society, Series B (Methodological) 55 (1993), 881-896.

[4] M. Ahdesmäki, H. Lähdesmäki, I. Shmulevich and O. Yli-Harja, Robust regression for periodicity detection in non-uniformly sampled time series, BMC Bioinformatics 6 (2013), pp. 117.