ROBUST DETECTION OF PERIODICALLY BEHAVING BIOLOGICAL TIME SERIES

Miika Ahdesmäki[1], *Harri Lähdesmäki*[1], *Ronald Pearson*[2], *Heikki Huttunen*[1] *and Olli Yli-Harja*[1]

[1]Institute of Signal Processing, Tampere University of Technology,
P.O.Box 553, FI-33101 Tampere, FINLAND, forename.surname@tut.fi,
[2]ProSanos Corporation, Harrisburg PA 17101, USA, ronald.pearson@prosanos.com

ABSTRACT

This paper describes a new way of assessing periodically behaving biological time series from multiple time series data, mainly aimed at studying microarray data. The method is based on robust spectral estimation and the g-statistic together with statistical significance value testing. The results show that this method performs extraordinarily well in processing both simulated and real microarray measurement data.

1. INTRODUCTION

Periodicity detection in time series measurements is a usual application of signal processing in studying biological data. Periodically behaving biological events can be of interest because of many reasons, e.g. periodicity in gene expression time series could be of interest because periodically behaving genes might suggest cell cycle control over the gene expression and so on. What is usually the case, is that there are lots of measured targets but very few time points per target. This is particularly true for gene expression measurements where there can be thousands of genes measured but only at few time points.

The task of finding periodicity in gene expression time series can be viewed as a decision problem based on spectral analysis. A formal statistical testing procedure for the detection of periodic expression profiles was recently introduced by [1]. It relies on the use of a so-called Fisher's g-statistic ([2]) for which the null-distribution can be derived under the Gaussian noise assumption. Because there are usually thousands of genes (time series) that are tested simultaneously, the obtained significance values must be corrected for multiple testing, using methods such as Bonferroni, step-down procedures, Benjamini and Hochberg or others (see, e.g., [3] for an extensive comparison).

A robust, rank-based, non-parametric spectral estimator was recently introduced in [4]. In this paper, we extend the approach of [4] to the detection of periodic time series, as introduced in [5]. This results in a robust testing procedure which is insensitive to a heavy contamination of outliers, missing-values, short time series, nonlinear distortions, and is completely insensitive to any monotone nonlinear distortions.

The rest of this paper proceeds as follows. The mathematical methods are first briefly introduced. Then we illustrate the power of the test under different noise conditions. As a conclusion, we discuss the properties of the presented methods.

2. MATHEMATICAL METHODS

Assume the model for a periodic time series as

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

where $\beta \geq 0, \omega \in (0, \pi), n = 1, \ldots, N, \phi \in (-\pi, \pi]$, and ϵ_n is an i.i.d. noise sequence. To test for the periodicity, define the null hypothesis as $H_0 : \beta = 0$, i.e., time series consists of the noise sequence alone, $y_n = \epsilon_n$.

2.1. The robust spectral estimator

We consider a recently introduced rank-based autocorrelation estimator [4] for the problems of spectrum estimation. This estimator is a moving-window extension of the Spearman rank correlation coefficient, quantifying the association between the sequences $\{y_k\}$ and $\{y_{k+m}\}$. The resulting quantity, $\rho^S(m)$ is actually an alternative estimator of the standard correlation coefficient $\rho(m)$ between these sequences. The estimator is defined as

$$\tilde{S}(\omega) = \sum_{k=-L}^{L} \tilde{\rho}(k) e^{-i\omega k}, \tag{2}$$

where $\tilde{\rho}(m)$ estimates the correlation coefficient between $\{y_k\}$ and $\{y_{k+m}\}$ and L is the maximum lag for which the correlation coefficient is computed. More specifically, we consider the correlation coefficient between the data ranks $R_y(i)$ and $R'_y(i)$ (let I_m be the set of time indices k for which both y_k and y_{k+m} are available and $K_m = |I_m|$), defined by

$$\rho^S(m) = \frac{1}{C} \cdot \frac{12}{K_m^2 - 1} \sum_{i \in I_m} \left(R_y(i) - \frac{K_m + 1}{2} \right) \cdot \left(R'_y(i) - \frac{K_m + 1}{2} \right), \tag{3}$$

where C is a normalisation factor, $R_y(i)$ denotes the rank of y_i in the set $S = \{y_j : j \in I_m\}$ and $R'_y(i)$ denotes the rank of y_{i+m} in the set $S' = \{y_{j+m} : j \in I_m\}$. By selecting either $C = K_m$ or $C = N$ in Eq. (3) yields the unbiased or the biased estimate of the correlation coefficient between the rank sequences, respectively. The biased version shall be used in this context due to its favorable properties.

2.2. The test statistic and significance values

After evaluating the spectral estimate of each time series in the multiple time series data, a significance value for each time series is evaluated based on a chosen statistic. Ideally a time series consisting of noise only gets a low significance value (a p-value close to one, indicating that the null hypothesis can not be rejected) and a time series having a strong periodic component gets a high significance value (p-value close to zero).

In the same way as Wichert *et al*. [1] do, we use the g-statistic and evaluate

$$g = \frac{\max_{1 \leq l \leq a} |\tilde{S}(\omega_l)|}{\sum_{l=1}^{a} |\tilde{S}(\omega_l)|} \tag{4}$$

for each time series spectral estimate. However, we do not have the luxury of resorting to an exact distribution of the g-statistic to find the p-values, e.g., under the Gaussian noise assumption. In some cases one might be interested in testing fixed instead of unknown frequencies. The proposed method can naturally be adapted to that case as well, i.e, in Eq. (4) we can choose some fixed power spectrum value instead of the maximum to test for that specific frequency.

To obtain the p-values we consider two common ways of computing them: simulation and permutation based methods. In the simulation approach we generate a set of, say, 10000 time series under the null hypothesis and evaluate the g-statistics of the time series in the set. Then we use e.g. kernel density estimation methods on the 10000 evaluated g-statistics to model the distribution. Then the computed g-statistics of the data of interest can be compared to the modelled distribution to find the p-values. In the permutation approach each time series is handled individually. For each time series we compute the g-statistic of e.g. 10000 (or some other number lower than $N!$, the maximum number of distinct permutations) distinct permutations of the time series at hand to model the g-statistic distribution for that specific time series. Using permutation tests in this way is of course much more computing intensive if compared to the simulation approach.

The obtained p-values are finally corrected for multiple testing by using the Benjamini and Hochberg method (false discovery rate, [3]).

3. POWER OF THE TEST

In this section we compare the power of the presented method to the classical periodogram approach since there is a strong connection between these two methods.

The power of the test, i.e., one minus the probability of the type II error (false negative), is estimated for three different test cases as well as for different time series lengths and for different noise parameters using 10000 Monte Carlo runs, see Figure 1. 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 shown in Eq. (1)

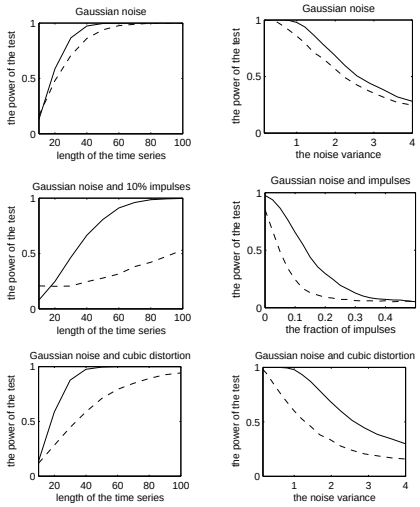

Figure 1. The power of the tests (y-axis) for the three different test cases as the function of the time series length and varying noise parameters (x-axis). The solid (resp. dashed) line corresponds to the proposed robust method (resp. Fisher's test).

with $\beta = \sqrt{2}$ to represent a periodic signal (i.e., the alternative hypothesis). In the right column of Figure 1, the length of the time series is set to 40 and the power is shown as the function of varying noise parameters. Figure 1 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 nonlinear distortion. More interestingly, however, the proposed method is also more powerful in the case of standard Gaussian noise.

4. CONCLUSION

As the results show, the presented robust method clearly outperforms analyses based on the classical periodogram.

5. REFERENCES

[1] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically series data," *Bioinformatics*, vol. 20, pp. 5 – 20, 2004.

[2] R. Fisher, "Tests of significance in harmonic analysis," *Proceedings of the Royal Society of London*, vol. 125, pp. 54–59, 1929.

[3] S. Dudoit, J. Shaffer, and J. Boldrick, "Multiple hypothesis testing in microarray experiments," *Statistical Science*, vol. 18, pp. 71–103, 2003.

[4] R. Pearson, H. Lähdesmäki, H. Huttunen, and O. Yli-Harja, "Detecting periodicity in nonideal datasets," *SIAM International Conference on Data Mining 2003, Cathedral Hill Hotel, San Francisco, CA, May 1-3*, 2003.

[5] M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja, "Robust detection of periodic time series measured from biological systems," *BMC Bioinformatics*, vol. 6:117, 2005.