ROBUST FISHER'S TEST FOR PERIODICITY DETECTION IN NOISY BIOLOGICAL TIME SERIES

Miika Ahdesmäki[1], *Harri Lähdesmäki*[1,2], *Olli Yli-Harja*[1]

[1]Institute of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland
[2] Institute for Systems Biology, WA 98103, USA,
miika.ahdesmaki@tut.fi, harri.lahdesmaki@tut.fi, olli.yli-harja@tut.fi

ABSTRACT

Periodicity detection in time series measurements is a usual application of signal processing in studying biological data. The reasons for detecting periodically behaving biological events are many, e.g. periodicity in gene expression time series could suggest cell cycle control over the gene expression. In this paper we present a robust version of the Fisher's test for detecting hidden periodicities in uniformly sampled time series data. The robust test performs better than the original test in case the data is not truly Gaussianly distributed. The proposed robust method is nearly as fast to evaluate as the original Fisher's test.

1. INTRODUCTION

Detecting periodically behaving biological time series has gained a lot of attention lately. Especially the growth of available gene microarray [1] time series data has led to the introduction of periodicity detection methods from other research fields of interest to gene expression studies.

Periodicity detection methods can be broadly divided into two classes. The generic approaches seek hidden periodic components at all the available frequencies [2, 3, 4, 5, 6] and use statistical tests to yield significance values with multiple correction. The methods in the second class seek periodic phenomena at a priori specified frequencies, e.g. the assumed cell cycle frequency, see e.g. [7, 8, 9, 10, 11, 12, 13].

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 [2] we presented a robust rank based modification of Fisher's g test [4] for finding hidden periodicities in time series data. 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 significance values.

In this paper we follow the general direction of Fisher's g test together with multiple testing correction for the detection of periodic time series in multiple time series data. We use a regression based formulation of the method presented [14] to find robust spectral estimates of time series instead of using the basic non-robust periodogram. After finding the spectral estimates we replace the periodogram in the g test with the robust spectral estimate, in a similar way as in [2].

To find the p values corresponding to the computed g statistics (the test statistic in Fisher's test), we propose to use the analytical distribution of the unmodified g statistic and explain the justification to do so. This way the robust test is nearly as computationally efficient as the original Fisher's test; thus no need for permutation tests or Monte Carlo simulations.

The performance of the proposed robust regression based method is then compared to that of the periodogram. We use simulations to evaluate the receiver operating characteristic (ROC) and power of test figures under several noise and signal configurations to show the performances.

2. METHODS

Fisher's test for the detection of hidden periodicities of unspecified frequency in time series ([4]) is a well known test based on the periodogram spectral estimator. The null hypothesis is that the time series is Gaussian noise against the alternative hypothesis that the signal contains an added deterministic periodic component of unspecified frequency. An implicit assumption is that the time series sampling is even. Thus, we 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 periodicity, define the null hypothesis as $H_0 : \beta = 0$, i.e., the time series consists of the noise sequence alone, $y_n = \epsilon_n$.

The g statistic in Fisher's test is then defined as

$$g = \frac{\max_{1 \leq i \leq q} I(\omega_i)}{\sum_{i=1}^{q} I(\omega_i)}, \tag{2}$$

with $q = [(n-1)/2]$ and $I(\omega)$ is the spectral estimate (the periodogram) evaluated at Fourier frequencies. The p

value for a realisation of the g statistic (g^*) is given by

$$P(g \geq g^*) = 1 - \sum_{j=0}^{q} (-1)^j \binom{q}{j} (1 - jg^*)_+^{q-1}, \quad (3)$$

and if this probability is less than α we can reject the null hypothesis at level α [7]. Fisher's test is further described in ([7]) and used successfully with gene expression microarray data in ([6]).

In case our measured time series are not strictly obeying the Gaussian distribution, but have for instance data points that are inconsistent with the rest of the data (outliers), we should consider making the test somehow more robust.

One way of gaining robustness is to use a robust time series cleaner and then use the Fisher's test with the cleaned time series. One such cleaner is introduced in ([14]) where the authors first estimate the Fourier coefficients of the time series by using robust M-estimation. The estimated robust Fourier coefficients are then inverse transformed with the ordinary inverse Fourier transform to yield a cleaned time series.

Instead of using the inverse Fourier transform as described and then using the periodogram in Fisher's test we drop the two unnecessary steps and use the originally M-estimated Fourier coefficients to form the robust periodogram. This way the Fourier coefficients are estimated directly with the Tukey's biweight-based M-estimator regression. By tuning the estimator correctly a 95% asymptotic efficiency on the standard normal distribution can be achieved. In addition, M-estimators are known to be asymptotically normally distributed. Since normally distributed Fourier coefficients form the basis of Fisher's classical result in Equation 3, asymptotic normality of M-estimators suggests a similar asymptotic null hypothesis distribution for the proposed robust test as well. This motivates us to use the original distribution of the g statistic (the statistic in Fisher's test) with our modified periodogram. Although the distribution is not exact for the modified test (see Figure 1), it is the exact distribution for the cleaned series.

Using 10000 Monte Carlo runs we estimated the null hypothesis distribution for the periodogram and for the modified estimator. For the modified estimator we used two null hypothesis signal types, one with Gaussian noise and the other with Gaussian noise and 10% outliers per time series. The time series length was set to 20 which is relatively short. The resulting distributional estimates and the theoretical distribution can be seen in Figure 1. Stronger tails are clearly visible for the modified test. However, as seen in Figure 2 (time series length 100), at longer time series lengths the distributions merge better.

When implementing the robust spectral estimator we first note that the periodogram is a function of the Fourier coefficients estimated from time series data. These coefficients can be seen as coordinates in a space whose basis is formed by orthogonal sinusoidals at the Fourier frequencies. We can thus represent the estimation of the Fourier

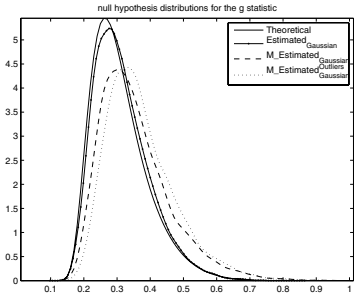

Figure 1. The ideal and estimated null hypothesis distributions at time series length 20.

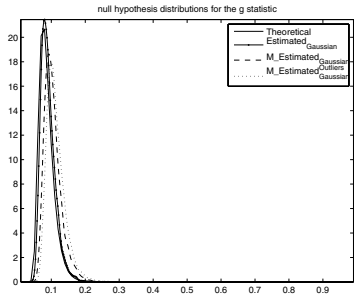

Figure 2. The ideal and estimated null hypothesis distributions at time series length 100.

coefficients as a linear model regression problem:

$$y = Xb, \quad (4)$$

The sinusoidals form the square model matrix X, y is the measured time series (or equivalently the coordinates in the time domain) and b are the Fourier coefficient to be estimated. When using least squares to solve the problem we can just invert X to get the estimate for b. This yields the Fourier coefficients for the periodogram. If we want to use robust regression for this we must reduce the dimensionality of the problem. Therefore we estimate the frequencies iteratively one at a time and always subtract the previously fitted part away before fitting the next frequency component since we no longer possess orthogonality properties. The order in which the frequencies are fitted is based on an initial spectral estimate where the frequencies are sought without subtracting the previously fitted parts and whose largest spectral component gives the first frequency to fit, the second largest the second frequency fit and so on. More details about the model can be found in ([15]) where we consider the more general case of non uniform sampling.

Based on the spectral estimates of the time series, we calculate the g statistics for each spectral estimate. A p value for each time series can then be found out with help of the distribution of the g statistic (Equation 3), telling us whether or not a strong periodic component is present. Multiple test correction of the p values is then necessary and we use the Benjamini-Hochberg [16] false discovery

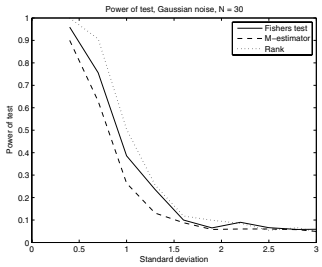

Figure 3. Power of test with varying Gaussian noise in the simulation data. The detectors are the basic Fisher's test (Fishers test), the M-estimator based Fisher's test (M-estimator) and the rank based estimator test ([2]Rank).

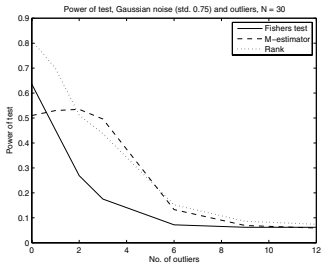

Figure 4. Power of test with varying amount of outliers in the simulation data.

rate (FDR) to choose a cut-off value for the p values accepted as periodic.

3. RESULTS

We now compare the power of the presented method to the basic Fisher's test and the method introduced in [2] since there is a strong connection between these methods. The power of the test, i.e., one minus the probability of the type II error (false negative), is estimated for two different noise parameter scenarios using 1000 Monte Carlo runs, see Figures 3 and 4. The significance level is set to $\alpha = 0.05$. In both the two 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) with $\beta = \sqrt{2}$ to represent a periodic signal (i.e., the alternative hypothesis). The lengths of the time series is set to 30 and the power is shown as the function of varying noise parameters.

Next we compare the receiver operating characteristics (ROCs) of the two methods. The ROC plots $sensitivity$ versus $1-specificity$ thus indicating the true positive rate while accepting more false positives. Figures 5–8 show 4 different scenarios. In Figure 5 there are no outliers present in the simulation data, in Figure 6 there is one outlier per time series and so forth. The main diagonal in these figures represents random decision; thus the name chance diagonal. The time series length in these simulations was 20.

As we can clearly see from the presented figures, a few outlying samples degrade the non-robust Fisher's test

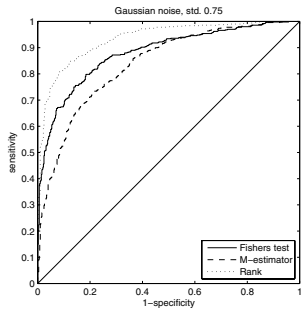

Figure 5. Receiver operating characteristic curves for the Fisher's test (Fishers test), the M-estimator based test (M-estimator) and the rank based estimator (rank[2]). The noise type in the time series is plain Gaussian with standard deviation 0.75.

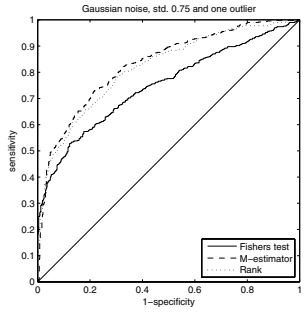

Figure 6. ROC curves with Gaussian noise and one outlier per time series.

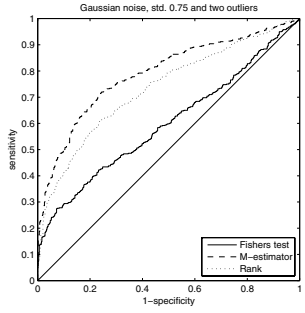

Figure 7. ROC curves with Gaussian noise and two outliers per time series.

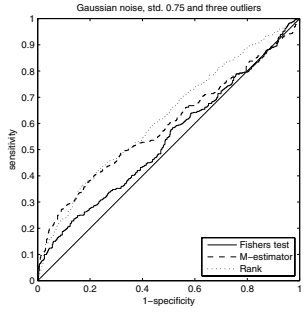

Figure 8. ROC curves with Gaussian noise and three outliers per time series

whereas the modified test can reject at least 10% of outliers. 10% of outliers in noisy short length data like cell level measurements is not too much considering the different sources of errors.

The results on real measurement data were shown to be very similar in [2] for both Fisher's test and the method presented in [2]. Since the new introduced method is very similar to these, the results are also likely to be similar and are not presented here due to limited space. Since the ground truth of measurement data is rarely known, the performance of different methods should always be verified by simulations.

4. CONCLUSION

As the results show, the presented robust method outperforms the unmodified Fisher's test when the data is not exactly normally distributed. Although the method presented in [2] outperforms the M-estimator modified test in most presented simulations (except in the 1 and 2 outlier cases), the implementation of the new test is almost as fast to evaluate as that of Fisher's test. If we neglect the effect of outliers and other deviations from the ideal model and use classical methods like Fisher's test we end up with biased results. Therefore we recommend to keep in mind when processing real measurement data that the classical methods pose a danger that should not be neglected.

5. ACKNOWLEDGMENTS

This work was supported by the Academy of Finland, project No. 213462 (Finnish Centre of Excellence program (2006 - 2011)). The support of Tampere Graduate School in Information Science and Engineering (TISE) is also gratefully acknowledged.

6. REFERENCES

[1] M. Schena, D. Shalon, R. Davis, and P. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, pp. 467–470, 1995.

[2] M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja, "Robust detection of periodic sequences in biological time series," *BMC Bioinformatics*, vol. 6, pp. 117, 2005.

[3] J. Chen, "Identification of significant genes in microarray gene expression data," *BMC Bioinformatics*, vol. 6, pp. 286, 2005.

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

[5] E. Glynn, J. Chen, and A. Mushegian, "Detecting periodic patterns in unevenly spaced gene expression time series using lomb-scargle periodograms," *Bioinformatics*, 2005.

[6] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically expressed transcripts in microarray time series data," *Bioinformatics*, vol. 20, pp. 5–20, 2004.

[7] P. Brockwell and R. Davis, *Time Series: Theory and Methods*, Springer-Verlag, New York, 2nd edition, 1991.

[8] U. de Lichtenberg, L. Jensen, A. Fausbø ll, T. Jensen, P. Bork, and S. Brunak, "Comparison of computational methods for the identification of cell cycle regulated genes," *Bioinformatics*, pp. (Advance Access published on October 28, 2004, doi:10.1093/bioinformatics/bti093), 2004.

[9] D. Johansson, P. Lindgren, and A. Berglund, "A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription," *Bioinformatics*, vol. 19, pp. 467–473, 2003.

[10] D. Liu, D. Umbach, S. Peddada, L. Li, P. Crockett, and C. Weinberg, "A random-periods model for expression of cell-cycle genes," *Proceedings of the National Academy of Sciences of the USA*, vol. 101, pp. 7240–7245, 2004.

[11] X. Lu, W. Zhang, Z. Qin, K. Kwast, and J. Liu, "Statistical resynchronization and bayesian detection of periodically expressed genes," *Nucleic Acids Research*, vol. 32, pp. 447–455, 2004.

[12] Y. Luan and H. Li, "Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data," *Bioinformatics*, vol. 20, pp. 332–339, 2004.

[13] L. Zhao, R. Prentice, and L. Breeden, "Statistical modeling of large microarray data sets to identify stimulusresponse profiles," *Proceedings of the National Academy of Sciences of the USA*, vol. 98, pp. 5631–5636, 2001.

[14] L. Tatum and C. Hurvich, "High breakdown methods of time series analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 55, pp. 881–896, 1993.

[15] M. Ahdesmäki, H. Lähdesmäki, I. Shmulevich, and O. Yli-Harja, "Robust regression for periodicity detection in non uniformly sampled time series," *Submitted to BMC Bioinformatics*.

[16] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc., Ser. B, Methodol.*, vol. 57, pp. 289–300, 1995.