Scientific Research and Essays Vol. 6(17), pp. 3717-3723, 26 August, 2011
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE11.802
ISSN 1992-2248 ©2011 Academic Journals

Full Length Research Paper

A filter based fisher g-test approach for periodicity detection in time series analysis

Masoud Yarmohammadi

Department of Statistics, Payame Noor University, 19395-4697 Tehran I. R. of Iran. E-mail: masyar@pnu.ac.ir.

Accepted 8 August, 2011

Periodicity is an interesting property of many time series data sets. A period can be defined as a self repeating pattern. This pattern provides useful information about the inherent structure in cyclic data set. In this paper, a filter based Fisher g-test approach is introduced. The filtering approach is based on the singular spectrum analysis. The power and running time of the proposed filter based approach are compared with non robust approaches. To evaluate the performance of the proposed approach we have performed a comprehensive simulation study. The results confirm the superiority of the proposed approach, considering various criteria which is insensitive to heavy contamination of outliers and short time series.

Key words: Periodicity, g-test, robustness, singular spectrum analysis.

INTRODUCTION

Periodicity provides useful information about the inherent structure in cyclic data set. For example the human respiration pattern is an example of an important periodic process. Deviation from normal periodic behavior is observed in many diseases. Periodicity can be used to derive the signature of normal breathing patterns and thereby facilitating abnormality detection. Periodicity not only helps to understand the properties of a single time series, but can capture complex relationships among multiple time series. For example, the heart rate, chest volume and blood oxygen concentration can be related through their periodic pattern. A fundamental nonparametric tool for detecting the periodicities of time series data is the periodogram. Although it is a basic spectrum estimation tool widely applicable in different application, but it is not a consistent estimator of the spectral density function. However, despite the inconsistency of the periodogram as a spectrum estimator, it is a useful tool for developing statistical inference methods for the spectral since its statistical properties are known. Consequently, many of the traditional statistical tests of the detection of periodic time series such as Fisher's test (Fisher, 1929) can be expressed in terms of the periodogram. Although the aforementioned methods provide exact test because they are based on a Gaussian assumption and a type of least squares estimation; they are not robust and can fail if the original noise assumptions do not hold. For example in many applications, the exact noise characteristics are usually unknown and can be remarkably non-Gaussian. Furthermore, the observed time series data can exhibit outliers, short length and distortion from the original wave form. Therefore, the computational methods should preferably be in robust such anomalies in the data. To solve this problem, a robust version of the fisher g-test has been introduced by Wichert et al. (2004) and Ahdesmaki et al. (2005). We review this method in this study and compare it with the proposed approach in there after. Here we consider another alternative approach.

According to this approach, we start with filtering the perturbed data in order to reduce the effect of existence of outliers and then we use fisher g-test. It is expected that the obtained results by this approach are more effective than the first two as we do not remove outliers. Furthermore, our proposed approach works very well even for a small sample size. Moreover, we reduce the noise level in order to increase the data quality improvement. In line with this research, it has been shown that noise reduction is important for curve fitting in the linear and nonlinear regression models (Hassani et al., 2009a, 2010a, b, c). The next challenge is to choose a proper filtering technique. There are several linear and

nonlinear methods for filtering noisy data. It has been shown that the singular value decomposition based techniques are more effective than the other ones for the noise reduction and filtering (Golyandina et al., 2001). Here, we use the singular spectrum analysis (SSA) technique which is an SVD-based approach as a filtering tool. SSA is designed to look for nonlinear, non–stationary, and intermittent or transient behaviour in an observed time series and following its successful application in the physical sciences, applications in economics and finance are now also finding favour (Thomakos et al., 2002; Hassani and Zhigljavsky, 2009; Hassani et al., 2009b). It is noticeable that there are several modification of SSA procedure (Golyandina et al., 2001; Hassani, 2010), however here we use basic version of SSA.

The structure of this paper is as follows: subsequently, we introduce the periodogram method as a standard periodicity detection tool to obtain Fisher's test and robust version of it, after which the new approach based on SSA is introduced. This is followed by a presentation of the result of comparison based on simulation studies. Finally, a brief conclusion is presented.

THE PERIODOGRAM AND FISHER G-TEST

Given a time series $[y_1,...,y_n]$ in the following Fourier representation:

$$y_t = \sum_{k=0}^{[n/2]} (a_k \cos w_k t + b_k \sin w_k t) + e_t \quad (1)$$

Where,

$$a_k = \begin{cases} \frac{1}{n}\sum_{t=1}^{n} y_t Cos\,\omega_k t \quad k = 0, \frac{n}{2} (if\ n\ is\ even) \\ \frac{1}{n}\sum_{t=1}^{n} y_t Sin\,\omega_k t \qquad k = 1,..., \left[\frac{n}{2}\right] \end{cases}$$

and

$$b_k = \frac{2}{n}\sum_{t=1}^{n} y_t Sin\,\omega_k t \qquad k = 1,..., \left[\frac{n-1}{2}\right]$$

Where e_t is the noise term with distribution $N(0,\sigma^2)$. In Model 1 we can test $H_0: a_k = b_k = 0 \quad vs \quad H_1: a_k \neq 0 \ or \ b_k \neq 0$. The fundamental, nonparametric tool for spectrum estimation is to use the periodogram as defined as follows:

$$I_k = I_n(\omega_k) = \frac{2}{n}\left|\sum_{t=1}^{n} y_t e^{-i\omega_k t}\right|^2 \quad (2)$$

We can also compute it at a discrete set of Fourier frequencies $\omega_k = 2\pi k / n,\ k = 0,1,...,[n/2]$. Preiestly (1981) showed that for each ω we may write $I_n(\omega)$ in an alternative form:

$$I_n(\omega) = 2\sum_{s=-(n-1)}^{(n-1)} \hat{R}(s)\cos(s\omega)$$

Where $\hat{R}(s) = \frac{1}{n}\sum_{t=1}^{n-|s|} y_t y_{t+|s|}$ is an autocovariance function at lag *s*. Thus, we are able to test whether a series contains multiple *m* periodic components by postulating the model. It should be noted that if we observe that the periodogram contains a number of sharp peaks, we should not conclude immediately that each of these peaks corresponds to a genuine periodic component obtained from series y_t. It has been recommended that we need to apply a suitable test to the periodogram peak to determine whether its value is significantly larger than that which would be likely to arise if there were no genuine periodic components in the model. The usual procedure is to start by plotting the periodogram ordinates at the standard frequencies $\omega_k = 2\pi k / N, k = 0,1,...[n/2]$, and then test the value of the largest observed peak. Fisher (1929) derived an exact test for the detection of hidden periodicities of unspecified frequency in time series based on the following statistic:

$$g = \frac{\max(I_k)}{\sum_{k=1}^{[n/2]} I_k} \quad (3)$$

Test statistic (3) known as Fisher's g-test statistic. Since the *g*-statistic divides the maximum periodogram ordinate by the sum of all periodogram ordinates, large values of *g* indicate a strong periodic component and can lead to the rejection of the null hypothesis. Fisher showed that (for the case *n* odd) the exact distribution of *g* under H_0 is given by:

$$p(g>z) = n(1-z)^{n-1} - \frac{n(n-1)}{2}(1-2z)^{n-1} + ... + (-1)^a \frac{n!}{a!(n-a)!}(1-az)^{n-1} \quad (4)$$

Where "a" is the largest integer less than $\left(\frac{1}{z}\right)$. Thus, for

any given significance level α, we can use Equation 3 to find the critical value g_α, such that $p(g > g_\alpha) = \alpha$. If the *g* value calculated from the series is larger than g_α then we reject the null hypothesis and conclude that the series y_t contains the specified periodic component (Wei, 1990).

Robust Fisher g-test

Let us turn back to the spectrum estimation problem. As it was mentioned in Priestley (1981), the periodogram $I_n(\omega)$ is equivalent to the correlogram spectral estimator $\hat{r}(k)$, that is:

$$\hat{f}(\omega) = \sum_{k=-L}^{L} r(k) e^{-i\omega k}$$

Where $\hat{r}(k)$ is the biased estimator of the autocorrelation function. Since the time series data is often contaminated with different types of outliers, the spectral estimation method and our test results are not reliable in most cases. To overcome this problem we consider a ranked based autocorrelation estimator for the problem of spectrum estimation. This estimator is a moving-window extension of the Spearman rank correlation coefficient, quantifying the association between sequences {yj} and {yj + k}. More specifically, we consider the correlation coefficient between the data ranks Ry(i) and $R'_y(i)$ defined by Ahdesmaki et al. (2005) as:

$$r^s(m) = \frac{1}{C}\frac{12}{k_m^2 - 1}\sum_{i \in I_m}\left(R_y(i) - \frac{k_m + 1}{2}\right)\left(R'_y(i) - \frac{k_m + 1}{2}\right) \quad (5)$$

Where C is a normalization factor, Ry(i) denotes the rank of yi in the set S = {yj : j ∈ Im} and $R'_y(i)$ denotes the rank of yi + m in the set S' = {yj + m : j ∈ Im}, where I_m is the set of time indices k for which both y_k and y_{k+m} are available and $k_m = |I_m|$. By selecting either C = Km or C = N in Equation 5 yields the unbiased or the biased estimate of the correlation coefficient between the rank sequences, respectively. The robust version of spectral density function is as follows:

$$\tilde{s}(\omega) = \sum_{k=-L}^{L} r^s(k) e^{-i\omega k}$$

Wichert et al. (2004) and Ahdesmaki et al. (2005) suggest using the *g*-statistic and evaluate the following statistic:

$$\tilde{g} = \frac{\max_{1 \le l \le a} |\tilde{s}(\omega_l)|}{\sum_{l=1}^{a} |\tilde{s}(\omega_l)|} \quad (6)$$

We called the aforementioned test statistic as 'robust fisher g-test statistic'. Note that, the exact distribution of the *g*-statistic, for example, under the Gaussian noise assumption is unknown. Therefore, to obtain the significance values we may consider the simulation studies. Moreover, this method requires intensive numerical computations.

FILTER BASED FISHER TEST

Here, we aim to use a filter based approach as a new approach to circumstances which there are outliers in the dataset. We use singular spectrum analysis (SSA) that is a powerful technique for time series analysis incorporating the elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing (Golyandina et al., 2001). In what follows we give a breif explanation of the SSA method (Hassani, 2007).

Singular spectrum analysis

The SSA technique consists of two complementary stages: decomposition and reconstruction and both of which include two separate steps. The original time series is decomposed into a number of additive time series, each of which can be easily identified as being part of the modulated signal or as being part of the random noise. This is followed by a reconstruction of the original series. A brief description of the method will be given here. Consider the real-valued non-zero time series $Y_T = (y_1, \ldots, y_T)$ of sufficient length T. Let $K = T - L + 1$, where L $(L \le T/2)$ is some integer called the 'window length'. Define the 'trajectory matrix':

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = [X_1 \ldots, X_K]$$

Where:

$$X_j = (y_j, \ldots, y_{L+j-1})^T.$$

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & \cdots & \vdots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_T \end{pmatrix} \quad (7)$$

Note that X is a Hankel matrix which means that all the elements along the off diagonal are equal. We then consider $\mathbf{X}$ as multivariate data with L characteristics and $K = T - L + 1$ observations. The columns $X_j = (y_j, \ldots, y_{L+j-1})^T$ of $\mathbf{X}$ considered as vectors lie in an L-dimensional space P^L. Define the matrix $\mathbf{X}\mathbf{X}^T$. Singular value decomposition (SVD) of $\mathbf{X}\mathbf{X}^T$ provides us with the collections of L eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L \geq 0$ and the corresponding eigenvectors $U_1, U_2, \ldots, U_L$, where U_i is the normalized eigenvector corresponding to the eigenvalue λ_i $(i = 1, \ldots, L)$. If we denote $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$ (i = 1, . . . , L), then the SVD of the trajectory matrix X can be written as:

$$\mathbf{X} = \mathbf{E_1} + \ldots + \mathbf{E_L}$$

Where:

$$\mathbf{E}_i = \sqrt{\lambda_i}\, U_i V_i^T$$

If we choose the first r eigenvectors $U_1, \ldots, U_r$, then the squared L_2-distance between this projection and $\mathbf{X}$ is equal to $\sum_{j=r+1}^{L} \lambda_j$. According to the basic SSA algorithm, the L-dimensional data is projected onto this r-dimensional subspace and the subsequent averaging over the off diagonals allows us to obtain an approximation to the original series. The main postulate of SSA procedure is that this approximation has the least noise effect, therefore we expected that the resluts obtained by this method have high precision.

SIMULATION STUDIES

Let us now evaluate the performance of our proposed approach using simulation study. The test signal model is as follows:

$$Y_t = \beta . \cos(\omega t) + \epsilon_t , \quad (8)$$

Where t = 1,…, *N* and ω is uniformly randomly chosen and ϵ_t is an i.i.d. noise sequence. We consider two types of noise levels:

i) Gaussian noise (zero mean).
ii) Gaussian noise and impulsive noise.

For the second case, we consider several data points randomly and multiply them with a constant number.

Power test

Let us now examine the power of our proposed test, that is, the probability that the test will reject a false null hypothesis. The power of the test is estimated for the three different procedures; that is, non-robust fisher g-test, robust fisher g-test and filter approach based on SSA. We have also considered different time series lengths and different noise parameters. The simulations were repeated 1000 times. The test power has been calculated as follows: using GenCycle package written by Ahdesmaki et al. (2005), we obtain the p-value of both the fisher-g-test and robust fisher g-test. Then, proportion of the rejection of false null hypothesis from 1000 p-value of the simulation runs gives the power test. Another point that we must to clarify is the parameters of the SSA. Certainly, the choice of the parameters depends on the data and the analysis we have to perform. Many rules have been proposed in the literature (Golyandina et al., 2001; Golyandina, 2010). According to common suggestion of the researchers for choosing the SSA parameters, we use half of the time series length for window length parameter. Choosing the number of needed singular values for the filtering stage, *r* depends on the structure of the series (Golyandina et al., 2001). Here we use *r* = 2 singular values to refine the series as we have a simple sine series and there is no intercept in the model. To gain a better understanding of the effect of filtering and evaluating the performance of the proposed approach, we consider our simulation studies with several α levels, β parameters, different percentage of contaminations and various noise levels. For all these the case-specific noise assumptions are used for both the null hypothesis (H_0: $\beta = 0$) and the alternative hypothesis (H_1: $\beta > 0$). Figures 1 to 4 represent the results. Solid, dashed and dotted lines denote SSA, non-robust fisher g-test and robust fisher g-test, respectively. Figure 1 shows the test power for periodicity detection with alpha levels 0.01, 0.05, 0.10 and 0.15 for Model 8 (considering $\beta = 2$ and $\omega = 0.05$). As can be seen from the figure, there is a significant difference between the power of the filter based approach and other approaches for all cases. Figure 2 shows comparison of three methods with selected parameters $\beta = 0.5, 1.0, 1.5$ and 2.0. The results confirm the superiority of filter based approach. Figure 3 shows the results for four noise levels $\sigma = 1.5, 2, 2.5$ and 3. As can be seen from the figure, the power of new approach becomes better by increasing noise level. Different percentages of contamination considered in Figure 4. The results indicate that the power of filtered based fisher g-test remains high in the circumstances of high percentages of contamination.

Running time

Let us now consider the performance of all aforementioned

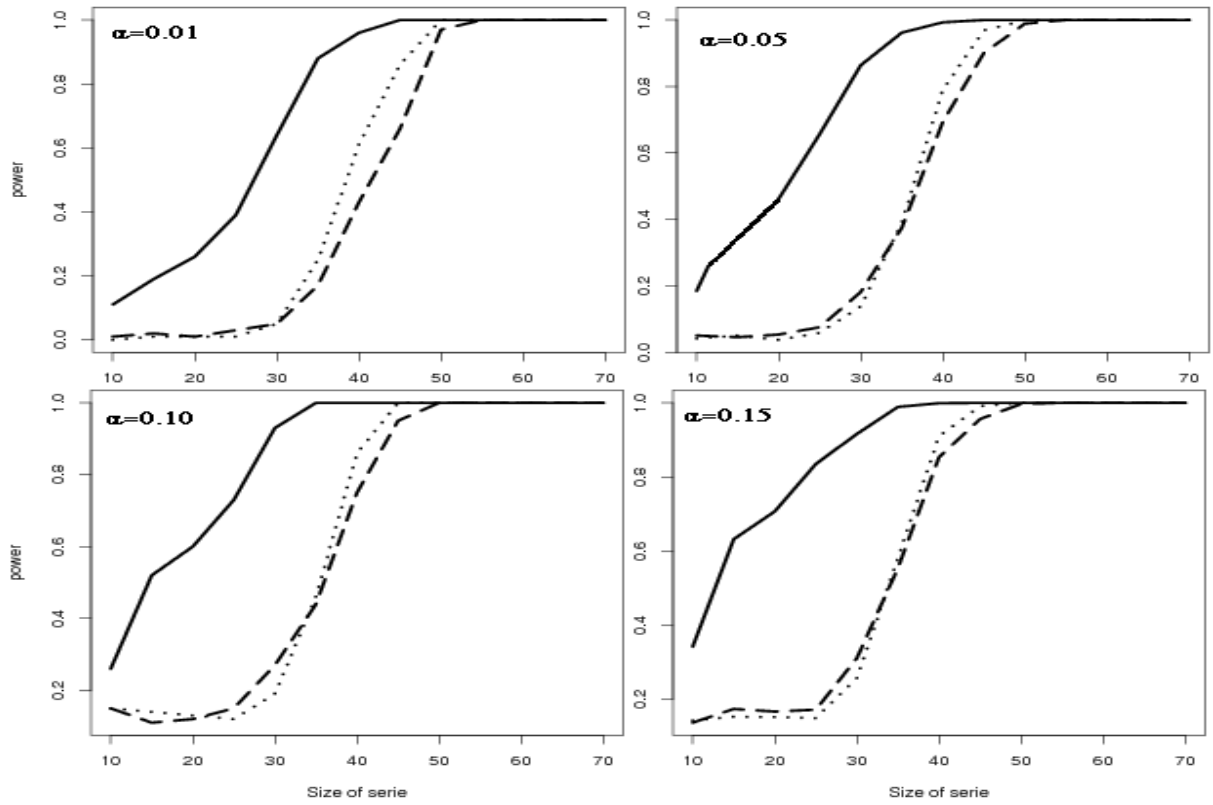

Figure 1. Power test with respect to Alpha.

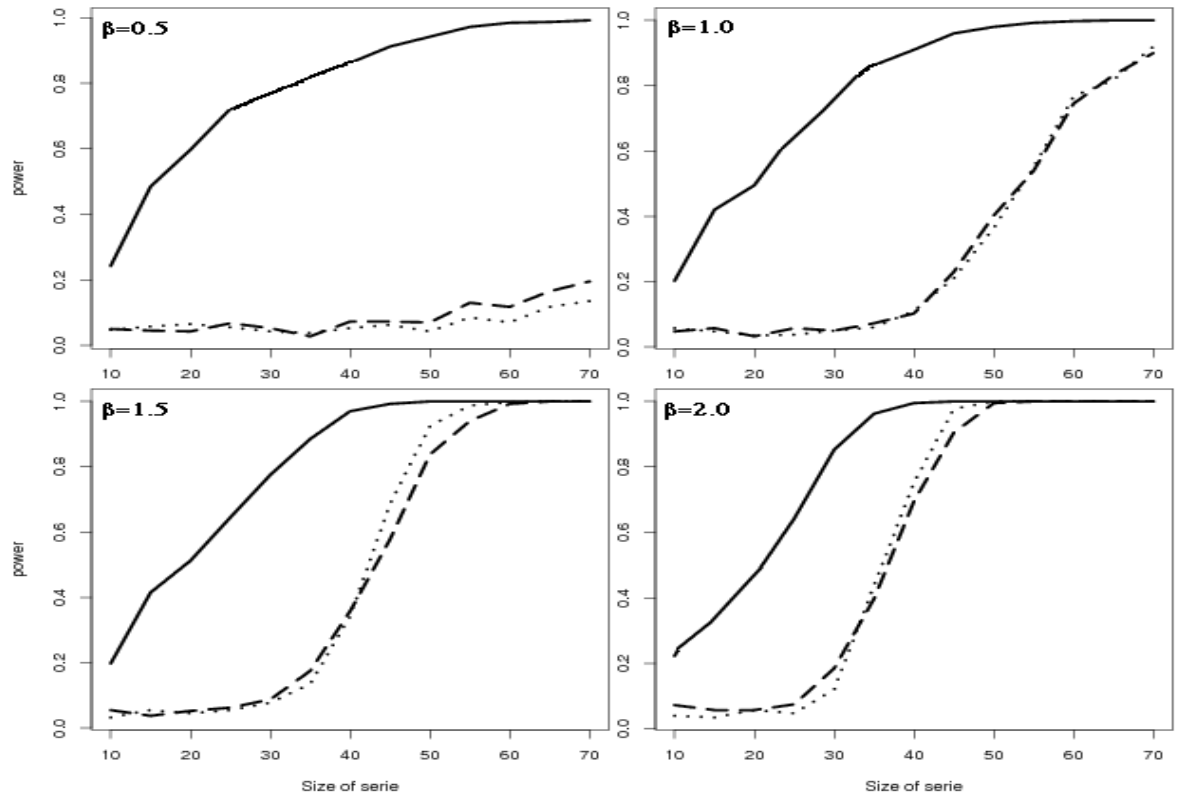

Figure 2. Power test with respect to parameter β.

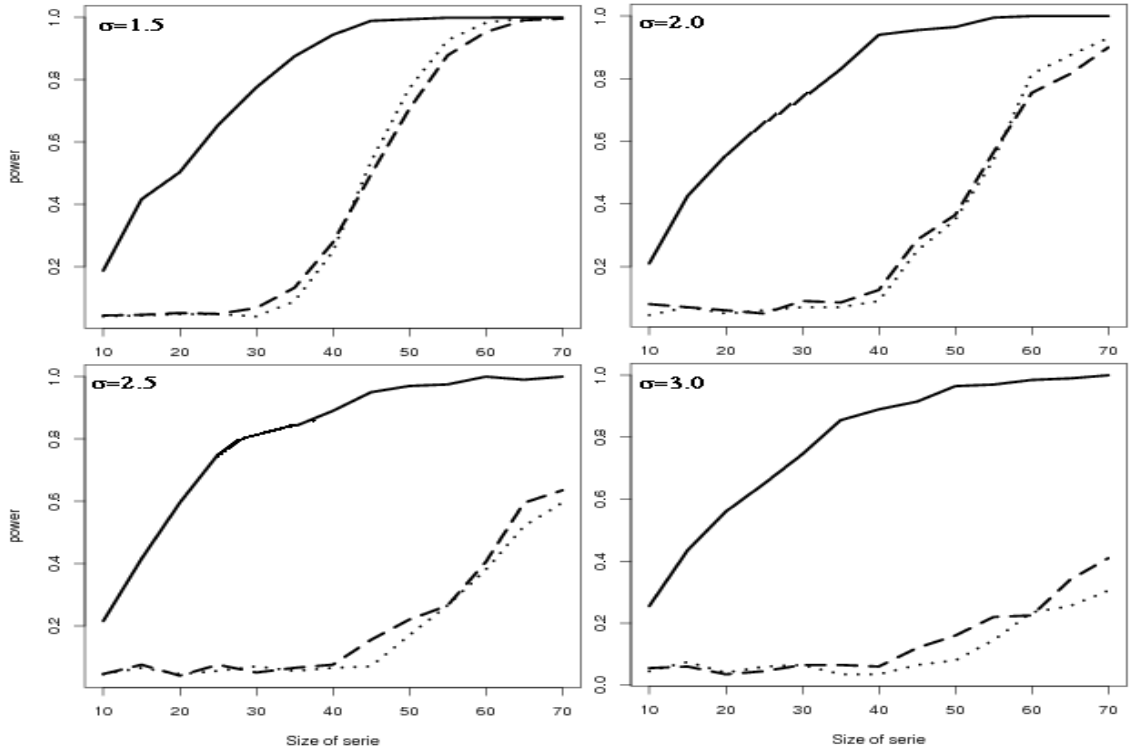

Figure 3. Power test with respect to noise level.

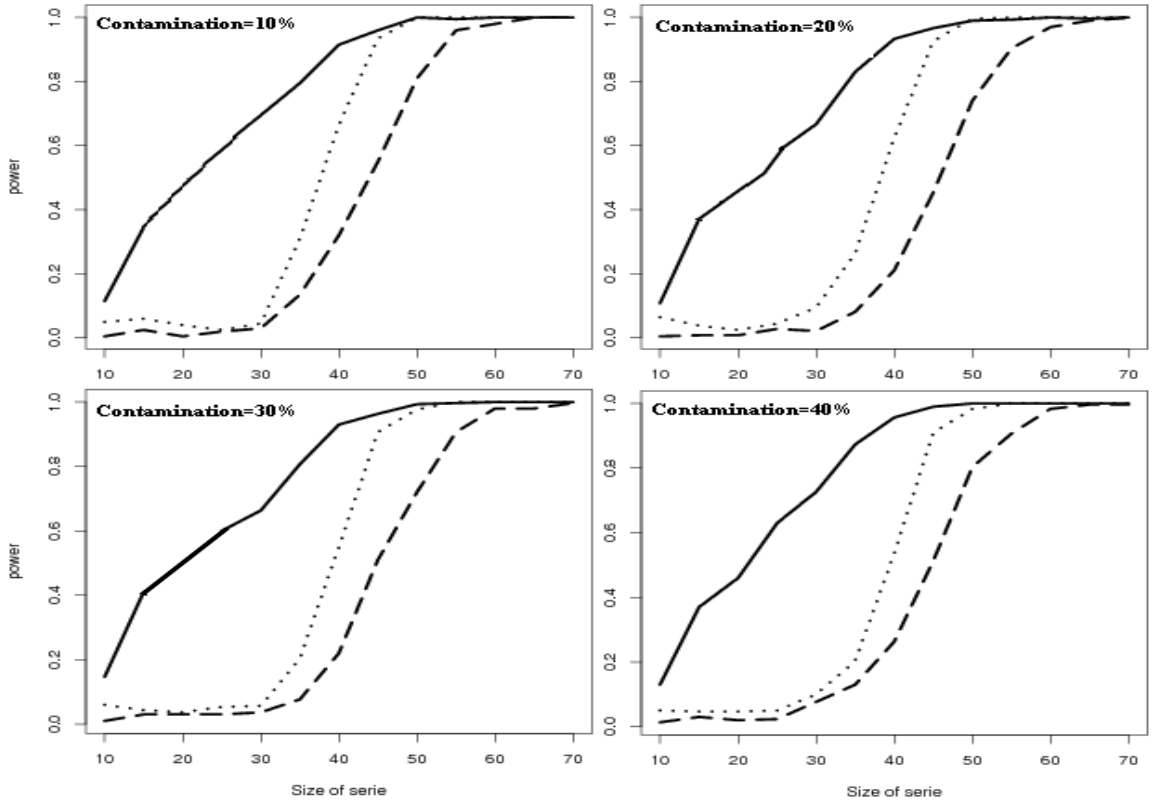

Figure 4. Power test with respect to contamination percent.

Table 1. Run time for three approaches. All programs were run on a computer with 2 GHz CPU and 2 GB of RAM.

Size of time series	Fisher g-test (s)	Filtered base Fisher g-test (s)	Robust fisher g-test (s)
10	0.00	0.00	53.07
20	0.00	0.00	109.25
50	0.00	0.00	285.92
100	0.00	0.02	590.07
200	0.00	0.06	1261.14

test with respect to running time. Table 1 represents the results. The results indicate that the running time of the proposed approach is also faster than the robust fisher g-test.

Conclusion

Our simulation results with strong evidence confirmed that the filtering approach using SSA and then using fisher g-test is more robust than the robust fisher g-test. The proposed approach yields powerful results in finding periodicity in time series. As illustrated in the simulations study, the proposed filtered based approach has clearly better performance than the Fisher test and robust version of it considering different aspects. Moreover, the results confirm that the running time of the filtered based approach substantially is less than the robust version and has been used so far.

ACKNOWLEDGMENT

This research was supported by a grant from Payame Noor University, Tehran-Iran.

REFERENCES

Ahdesmaki M, Lähdfdmäki H, Pearson R, Huttenen H, Yli-Harja O (2005). Robust detection of periodic sequences in biological time series. BMC Bioinforma., 6: 117.

Fisher RA (1929) Tests of Significance in Harmonic Analysis. Proc. Roy. Soc. Ser. A., 125: 54-59.

Golyandina N, Nekrutkin V, Zhigljavsky A (2001). Analysis of Time Series Structure: SSA and Related Techniques. Chapman & Hall/CRC, New York, London.

Golyandina N (2010). On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods. Statistics and Its Interface, 3(3): 257-279.

Hassani H (2007). Singular Spectrum Analysis: Methodology and Comparison. J. Data Sci., 5(2): 239-257.

Hassani H, Zokaei M, Von Rosen D, Amiri S, Ghodsi M (2009a). Does Noise Reduction Matter for Curve Fitting in Growth Curve Models? Comput. Methods Program Biomed., 96(3): 173-181.

Hassani H, Heravi S, Zhigljavsky A (2009b). Forecasting European Industrial Production with Singular Spectrum Analysis. Int. J. Forecast., 25: 103-118.

Hassani H, Zhigljavsky A (2009). Singular Spectrum Analysis: Methodology and Application to Economics Data. J. Syst. Sci. Complexity, 22(3): 372-394.

Hassani H, Thomakos D (2010a). A Review on Singular Spectrum Analysis for Economic and Financial Time Series. Statistics and Its Interface,3(3): 377-397.

Hassani H, Mahmoudvand R, Yarmohammadi M (2010b). Filtering and Denoising in Linear Regression Analysis. Fluctuation and Noise Letters, 9(4): 343-358.

Hassani H, Dionisio A, Ghodsi M (2010c). The effect of noise reduction in measuring the linear and nonlinear dependency of financial markets. Nonlinear Analysis: Real World Appl., 11(1): 492-502.

Hassani H (2010). Singular Spectrum Analysis Based on the Minimum Variance Estimator. Nonlinear Analysis: Real World Appl., 11(3): 2065-2077.

Priestley MB (1981). Spectral Analysis and Time Series. Academic Press.

Thomakos DD, Wang T, Wille LT (2002). Modeling Daily Realized Futures Volatility with Singular Spectrum Analysis. Physica A: Stat. Mech. Appl., 312(3-4): 505-519.

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

Wei WWS (1990). Time Series Analysis: Univariate and Multivariate Methods. Addison Wesley, 1st Edition.