Research Article

Use of SSA and MCSSA in the Analysis of Cardiac RR Time Series

R. A. Thuraisingham

1A, Russell Street, Eastwood, NSW 2122, Australia

Correspondence should be addressed to R. A. Thuraisingham; ranjit@optusnet.com.au

Received 18 October 2013; Revised 23 November 2013; Accepted 23 November 2013

Academic Editor: Gabriela Mustata Wilson

Copyright © 2013 R. A. Thuraisingham. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A new preprocessing procedure in the analysis of cardiac RR interval time series is described. It uses the singular spectrum analysis (SSA) and the Monte Carlo SSA (MCSSA) test. A novel feature of this preprocessing procedure is the ability to identify the noise component present in the series with a given probability and to separate the time series into a trend, signal, and noise. The MCSSA test involves testing whether the modes obtained from SSA can be generated by a noise process leading to separation of the noise modes from the signal. The procedure described here does not discard or modify any sample in the record but merely separates the time series into a trend, signal, and noise, allowing for further analysis of these components. The procedure is not limited to the length of the record and could be applied to nonstationary data. The basis functions used in SSA are data adaptive in that they are not chosen a priori but instead depend on the data set used, increasing flexibility to the analysis. The procedure is illustrated using the RR interval time series of a healthy, congestive heart failure, and atrial fibrillation subject.

1. Introduction

Singular spectrum analysis (SSA) [1–4] is an analytical tool that is used in time series analysis. Although it has been used widely in the analysis of environmental data [1–6] its use in biomedical signals has not received much attention. Some of biomedical signals that have been used to illustrate SSA are electroencephalogram (EEG) signals collected during evoked potential studies [7], ultrasound [8], and the electrocardiogram signal (ECG) [9, 10]. In the latter it was used to separate the fetal ECG signal from the maternal signal and in the separation of the heart sound artifact from the respiratory signals. This paper is an attempt to illustrate its value in the analysis of the cardiac RR time series data. The main focus in this paper is on preprocessing the cardiac RR interval series. It proposes an alternate procedure to the use of wavelet analysis for trend removal and the use of impulse rejection filter to remove artifacts [11]. Measured cardiac data contains a large amount of noise and nonlinear trend which require separation before any statistical assessment on the signal can be carried out. The method proposed here uses SSA, followed by the Monte Carlo singular spectrum analysis test (MCSSA) [1, 12]. It is a novel technique not only to separate the signal into various SSA modes but also to identify the SSA modes which correspond to noise. The MCSSA test [1, 12] involves testing the SSA modes of the original data with a large number of surrogate noise time series generated from random data having similar statistical properties of variance and autocovariance of lag 1 as the original series. The test identifies with a certain probability that certain SSA modes could be associated with autocorrelated noise. The use of such a test helps in the separation of the raw time series into a signal and noise. Some of the components in the signal are then fitted to the raw time series to match the trend, separating the raw series further into a trend, signal, and noise. The procedure is not limited to stationary data. The ability to identify certain SSA modes with noise processes of a given probability is a novel feature of the proposed procedure. This provides a certain confidence in the identification of the signal. Further the basis functions used in SSA for the separation of the signal are not chosen a priori as in the wavelet procedure but instead are dependent on the data used.

The SSA method and the MCSSA are described in Section 2. Section 3 is the results of the application of this technique in the analysis of three RR interval time series one
from a healthy subject \((N)\), the other two from subjects having congestive heart failure (CHF) and atrial fibrillation (AF). The data was obtained from MIT-BIH Normal Sinus Rhythm database and BIDMC Congestive Heart Failure database and MIT-BIH Atrial Fibrillation database, posted on Physionet [13]. Section 4 is the conclusion.

2. Method

2.1. Singular Spectrum Analysis. The term singular spectrum analysis comes from the spectral (eigenvalue) decomposition of a matrix \(X\) into its set (spectrum) of eigenvalues. These eigenvalues \(\lambda\) are the numbers that make the matrix \(X - \lambda I\) singular where \(I\) is the unit matrix. The singular spectrum analysis (SSA) is the analysis of the time series using the singular spectrum. It was first introduced by Fraedrich [2] and Broomhead and King [3]. An excellent exposition of this technique is given by Elsner and Tsonis [1].

The embedding dimension or also known as the window of size \(N\) is done by making lagged or delayed copies of a segment of the time series. Suppose there are \(N\) values of a uniformly sampled time series \(x_i, i = 1, 2, \ldots, N\). An embedding dimension \(m\) is then chosen. The data is then arranged into an \((N \times m)\) matrix \(X\) with \(N = N_t - m + 1\) as

\[
X = \frac{1}{\sqrt{N}} \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_m \\ x_2 & x_3 & x_4 & \cdots & x_{m+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_N & x_{N+1} & x_{N+2} & \cdots & x_{N+m-1} \end{pmatrix},
\]

where \(\sqrt{N}\) is a convenient normalization. The constructed matrix \(X\) is called the trajectory matrix, and it contains the complete record of patterns that have occurred within a window of size \(m\). The elements of the matrix \(X\) are constructed as

\[
X(i, j) = x_{i+j-1}, \quad 1 \leq i \leq N = N_t - m + 1; \quad 1 \leq j \leq m.
\]

The embedding dimension or also known as the window width determines the longest periodicity captured by SSA. A lagged covariance matrix \(S\) is then computed as the product of the trajectory matrix \(X\) and its transpose \(X^T\) given by

\[
S = X^T X.
\]

The matrix \(S\) is a real square symmetric \((m \times m)\) matrix. Such a matrix can be made diagonal using a matrix \(E\) whose columns are orthonormal and a diagonal matrix \(\Lambda\) such that

\[
S = E \Lambda E^T.
\]

The square roots of the ordered real eigenvalues \(\lambda_k\), \(k = 1, \ldots, m\), of the matrix \(S\) are collectively called the singular spectrum. The corresponding eigenvectors \(E(\cdot, k)\) are referred to by different names such as basis functions, eigen vectors, and empirical orthogonal Functions (EOF). \(E^T\) is the transpose of \(E\).

The principal component \(a_k\) or sometimes referred to as the mode has \(N\) components where each component is obtained by projecting a time record each with \(m\) components onto to \(m\) components of the eigenvector \(k\). The principal component matrix will therefore be an \(m \times N\) matrix, where each row corresponds to an eigenvector with the \(N\) columns for the \(N\) time lags. The element \(a(k, i)\) is given by

\[
a(k, i) = \sum_{j=1}^{m} x_{i+j-1} E(j, k), \quad i = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, m.
\]

(5)

The eigen values \(\lambda_k\) correspond to the variance of each SSA mode \(a(k)\).

As a result of orthogonality, each mode can be probed independently from the remainder of the time record. The entire time series can be reconstructed via reconstructed components (RCs) where each RC is associated with a mode and EOF. The component \(RC^k\) can be written as

\[
RC^k_{i+1} = a(k, i) E(j, k), \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, m.
\]

(6)

The original signal can be reconstructed from the RC’s via

\[
x_{i+1} = \sum_{k=1}^{m} RC^k_{i+1}, \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, m.
\]

(7)

If filtering is carried out in the above summation \(k\) is restricted to values less than \(m\).

Once the eigenvalues \(\lambda_k\) of the lagged covariance matrix \(S\) are determined, a plot of \(\lambda_k\) as a function of the mode \(k\) is carried out. If white noise is present there will be a clear break in the singular values, with the eigenvalues associated with the noise spreading out almost flat with low values. On the other hand if the noise is red there will be a gradual exponential decay which will make it difficult to separate the signal from noise. In the next section identifying the modes related to noise and signal will be carried out using MCSSA.

An oscillatory mode is characterized by a pair of nearly equal SSA eigen values and associated modes that are in approximate phase quadrature [1, 4]. Such a pair can represent efficiently a nonlinear, anharmonic oscillation. This allows an oscillatory mode to be detected. A single pair of data-adaptive SSA EOF’s often will capture the basic periodicity of an oscillatory mode better than methods with fixed basis functions such as sines or cosines used in Fourier transform.

2.2. MCSSA. Monte Carlo SSA (MCSSA) [1, 12] is used to establish whether a given time series is distinguishable from a linear stochastic process which is normally considered as noise. Noise processes can be a described by a first order autoregressive process (AR(1)) where the value at a time \(t(\tau)\) depends on the value at time \(t = 1(\tau_{-1})\) only

\[
u_t = c u_{t-1} + e_t,
\]

(8)
where \( c \) is a constant that represents the lag correlation between successive observations in the time series, referred to as the AR(1) coefficient, and \( e_t \) is a random error term with mean zero and variance \( \sigma^2 \). If \(|c| < 1\), the series is stationary. If the noise is white, \( c = 0 \). For autocorrelated noise such as red noise where the power declines monotonically with increasing frequency, the range of \( c \) is within \( 1 > c > 0 \).

For a series containing only white noise the eigen values of the lagged covariance matrix \( S \) are all equal. Further for a time series which contains a signal bearing component which is contaminated with white noise, the eigenvectors of the lagged covariance matrix \( S \) and \( S^{\text{signal}} \) are the same, and \( \lambda_k = \lambda_k^{\text{signal}} + \nu^2 \), where the noise variance is \( \nu^2 \) [1, page 71].

In order to determine whether the type of eigen spectrum obtained from the measured time series can be generated by an autocorrelated noise process, one estimates the AR(1) coefficient and the variance of the random error series \( e_t \) from the measured data, assuming it to be an AR(1) process. An estimate of the AR(1) is obtained from the normalized autocovariance coefficient (\( c \)) of the measured data at lag 1 and the variance of the error series \( \sigma^2 \) from the variance of the data (\( \sigma^2 \)) and \( c, \sigma^2 = (1-c^2) \sigma^2 \) [14]. A number of different noise surrogate series \( \{u_t\} \) are then generated to identify the auto correlated noise process as an AR(1) process of the type shown by (8). They are obtained with \( u_0 = 0 \), the AR(1) coefficient \( c \) and with different realizations of random data all having the same variance \( \sigma^2 \). For each realization of the surrogate noise series the lagged covariance matrix \( S^{\text{noise}} \) is evaluated and the eigenvalues \( \{\lambda_k^{\text{noise}}\} \) are obtained from the diagonal elements of \( \Lambda^{\text{noise}} \) where \( \Lambda^{\text{noise}} = E_X S^{\text{noise}} E_X^T \). An identical transformation is applied to each surrogate \( S^{\text{noise}} \) to obtain the eigen values \( \{\lambda_k^{\text{noise}}\}, k = 1, 2, \ldots, m \). There is no need to diagonalize each \( S^{\text{noise}} \). The procedure measures the resemblance of a given surrogate with the measured data set. The ensemble of the sets of eigenvalues from the surrogate noise series is then used to test the hypothesis whether the eigen value of the measured time series can be attributed to an oscillatory mode. The 95 percentile of \( \{\lambda_k^{\text{noise}}(j)\} \) \( j = 1, 2, \ldots, n_s \) where \( n_s \) is the number of surrogates are then computed for each SSA mode \( k \). If the eigen value \( \lambda_k \) of the measured time series is greater than the corresponding 95 percentile value for the eigenvalue of the surrogate data then the probability that the SSA mode is due to autocorrelated noise is less than 5%.

### 3. Results and Discussion

In this section the above procedure is illustrated for three RR data sets obtained from N, CHF, and AF subject [13]. These three RR series have different characteristics and will provide a good example to understand this technique. The study in this report is limited to these three series and not to a bigger group since the focus of this study is to illustrate the technique instead of a statistical study on a group of RR series using this technique. In each of these cases 500 RR intervals were used. The embedding dimension \( m \) was chosen to have a value of 248. This choice was based on theoretical results and studies based on a wide class of simulated and real data which indicate that the optimal choice of the embedding dimension or window length is close to one half of the time series length \( N \) [15, 16]. The number of noise surrogates used in MCSSA was 1000. That is, the surrogate list consists of 1000 different noise time series, with each time series containing 500 samples, the same number of samples as the time series that is being studied. This choice for the number of surrogates provided significance estimate to an accuracy of 1%, that is impossible to differentiate between 95% and 95.5% [1, page 97].
Figure 1 shows the eigenvalues of the SSA modes of the data (red) along with the 95 percentile eigenvalues from an ensemble of 1000 surrogates generated by an AR(1) noisy process (blue). The information necessary to generate the noisy data is obtained from the measured data as described in Section 2.2. The SSA modes which are above the 95 percentile line are given in Table 1. These numbers total 31. That is, there are 31 eigenvalues associated with the signal and 217 with noise. Of this number, the modes where the consecutive eigenvalues are equal correspond to oscillatory modes having the same frequency but with phases offset by $\pi/2$. They are given in the second column of Table 1.

Before SSA is carried out the linear trend along with any residual mean is removed. The linear trend is removed by finding the best fit to a linear polynomial. The coarse nonlinear trend present in the series is then estimated from the first few signal modes with very high eigen values compared to the rest. In this case the first two SSA modes contribute to this coarse nonlinear trend. The sum of the RC’s that correspond to this nonlinear trend (SSA modes 1 and 2) is then added to the linear trend and mean to construct the observed trend of the raw RR series. This is shown as a red line superimposed on the raw RR series in Figure 2(a), where the healthy RR intervals are drawn. Figure 2(b) shows the sum of the RC’s that constitute the signal SSA modes without the RC’s of the trend. Figure 2(c) shows the part that is attributed to noise.

Figures 3 and 4 are similar plots for the CHF data while Figures 5 and 6 correspond to AF data. The numbers of the SSA modes that correspond to the signal and to the periodic signals are given in Table 1, both for the CHF and AF data.

The above results for the three different RR interval data sets illustrate the ability of the SSA technique to separate the raw series into a trend, signal, and noise. The trend seen in the raw time series is matched by the linear trend and a few low frequency RC’s from the SSA. These RC’s correspond to SSA modes having high eigen values, in comparison to the rest. Matching of the observed trend is carried out by successive inclusion of the RC’s to the linear trend, beginning with the RC that corresponds to the highest eigen value, until a reasonable smooth reproduction of the trend is seen visually. The RC’s that are used to construct the trend are a small subset of the RC’s that are identified as the signal and obtained after filtering the noise components. For the examples chosen in this paper, the number of RC’s that were used to construct the trend corresponded to the first two eigen values for N, first four for the CHF and the first eleven for the AF data. The eigen values corresponding to these RC’s are quite separate from the rest. Although in this paper the RC’s are chosen to visually match the trend observed, it is not limited to this procedure. Any other quantitative method could be chosen but it was found that the visual criterion was the safest procedure. What is identified is the coarse trend, and any procedure chosen must be verified visually to ensure that it is the coarse trend that is removed and not the signal.

The method adopted here to construct the trend and the signal is an alternate procedure to the use of wavelet analysis and the use of impulse rejection filter [11]. In the procedure mentioned in reference [11], the wavelet procedure...
Figure 4: CHF subject: (a) raw RR series (blue) and the trend (red); (b) the signal component of the raw series without the trend; (c) component of the raw series attributed to noise.

Figure 5: Eigen values of the SSA modes of the AF data (red) along with the 95 percentile eigenvalues from an ensemble of 1000 surrogates generated by an AR(1) noisy process (blue).

is used to remove the trend and the impulse rejection filter to remove the RR ectopic beats. The removed ectopic beats are replaced with the median RR value found in the adjacent sectors. Thus there is replacement of the original signal and no identification of the presence of colored or white noise. The present procedure is therefore different and superior. There is no replacement of any part of the signal but merely a separation of the signal into three categories: trend, signal, and noise. Further one of the main advantages of the present method is the filtering process for the noise which is verified by MCSSA.

The standard deviations of the signal for the N, CHF, and AF data are 0.067, 0.012, and 0.150 s, respectively. The depressed value for the standard deviation of the signal in CHF compared to N is in agreement with previous studies [11, 17]. Although the multiscale entropy (MSE) of a long AF record indicates that it can be approximated as white noise [18], the results on the SSA analysis done on a subset 500 samples from this same record indicate that such an approximation is invalid. If the AF record was white noise then the eigenvalues of SSA should all be equal [1, page 70]. This is not the case as can be seen from Figure 5 where the eigenvalues of AF are plotted. Further if the AF record is generated by a noise process, in MCSSA all the eigenvalues of the AF record should have been below the 95 percentile line, which is not evident. Also the visual presence of a nonlinear trend seen in the AF record is further evidence that there is correlation from neighboring samples. Such a result from a subset of a long record is not unexpected since the statistical properties of a finite sample from a large population will not always reflect the statistical properties of the larger group. This is the case when there are differing local time variations, which are averaged out when the sample number is increased. However the analysis of AF shows characteristics which indicate that it is more random than those of the CHF and N record. Let us examine this via the evaluation of the AR(1) coefficient which gives the correlation amongst neighboring samples.

Although the noise component is identified using the MC procedure, further separation of this noise as external or due to some internal cardiac process is not feasible at this stage. Assuming that all noise is internal, the AR(1) coefficient evaluated after removing only the linear trend from the data gives values of 0.493, 0.145, and −0.099 for the N, CHF and AF records, respectively. This shows that the correlation amongst the neighboring samples is the greatest for the N record with diminished correlations for the diseased subjects. Correlation is the least for the AF record indicating that it is more random. Since we do not expect all noise to be internal, another evaluation is carried out in the limit where all noise is external. In this case the linear trend and all noise are removed from the data sets and the AR(1) coefficient evaluated. The values for the AR(1) coefficient for the N, CHF, and AF records are 0.621, 0.231, and −0.101, respectively.
Again the neighboring correlations are in the order N > CHF > AF, with the AF record exhibiting the most random behavior. The presence of strong correlations in the healthy record and the diminished correlations amongst the diseased subjects is in agreement with the results of Costa et al. [18] showing that RR interval times of a healthy subject is more complex compared to those of a diseased subject.

The procedure outlined above using SSA does not discard or modify any samples in the record. It merely separates into different categories of nonlinear trend, signal, and noise. The nonlinear trend comprises a part the low frequency signal component which is added to the linear trend to fit the observed trend in to the measured data. The noise component is identified using a Monte Carlo procedure. Such an identification procedure for the noise component is an important feature of SSA. Since no part of the raw series is modified, with further available knowledge of external noise sources, it may be possible to separate noise components that are external from those that are heart related. This however is not carried here, since information about the external noise is unavailable. Although the example used in the illustration here has used a moderately small sample size, the method is not limited by such considerations. The choice of a moderate sample was made here merely from limitations to present details of such data in text. The division of a large sample into moderate subsets and performing SSA on them and then combining the results of the subsets is a computationally efficient procedure that can be adopted for large data sets.

4. Conclusion

A new preprocessing procedure in the analysis of cardiac RR interval time series using SSA is described. A Monte Carlo test is further carried to identify the noise component present in the time series. The latter procedure involves testing whether the SSA modes can be generated by a noise process, leading to a separation of the noise modes from the signal. Such an identification technique for the noise component is a novel feature of this preprocessing procedure. The procedure does not discard or modify any sample in the record but merely separates the time series into a trend, signal, and noise, allowing further analysis of these components. The procedure is not limited to the length of the record and could be applied to nonstationary data. Another feature of this procedure is the basis functions used in the SSA analysis which are not chosen a priori as in wavelet analysis but instead are dependent on the data set used. The method adopted here to construct the trend and the signal is an alternate procedure to the use of wavelet analysis and the use of impulse rejection filter [11]. The method is illustrated using as an example RR interval time series of a healthy, congestive heart failure, and atrial fibrillation subject.

References

[1] J. B. Elsner and A. A. Tsonis, *Singular Spectrum Analysis: A New Tool in Time Series Analysis*, Plenum Press, New York, NY, USA, 1996.

[2] K. Fraedrich, “Estimating the dimensions of weather and climate attractors,” *Journal of the Atmospheric Sciences*, vol. 43, pp. 419–432, 1986.

[3] D. S. Broomhead and G. P. King, “Extracting qualitative dynamics from experimental data,” *Physica D*, vol. 20, no. 2-3, pp. 217–236, 1986.

[4] M. Ghil, M. R. Allen, M. D. Dettinger et al., “Advanced spectral methods for climatic time series,” *Reviews of Geophysics*, vol. 40, no. 1, pp. 3–1–3–41, 2002.

[5] R. Vautard and M. Ghil, “Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series,” *Physica D*, vol. 35, no. 3, pp. 395–424, 1989.

[6] H. N. Southgate, K. M. Wijnberg, M. Larson, M. Capobianco, and H. Jansen, “Analysis of field data of coastal morphological evolution over yearly and decadal timescales. Part 2: non-linear techniques,” *Journal of Coastal Research*, vol. 19, no. 4, pp. 776–789, 2003.

[7] A. M. Tome, A. R. Teixeira, N. Figueiredo, I. M. Santos, P. Georgieva, and E. W. Lang, “SSA of biomedical signals: a linear invariant systems approach,” *Statistics and Its Interface*, vol. 3, no. 3, pp. 345–355, 2010.
[8] W. C. A. Pereira, S. L. Bridal, A. Coron, and P. Laugier, “Singular spectrum analysis applied to backscattered ultrasound signals from in vitro human cancellous bone specimens,” IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 51, no. 3, pp. 302–312, 2004.

[9] M. Ghodsi, H. Hassani, and S. Sanei, “Extracting fetal heart signal from noisy maternal ECG by multivariate singular spectrum analysis,” Statistics and Its Interface, vol. 3, no. 3, pp. 399–411, 2010.

[10] S. Sebastian and S. Rathnakara, “Separation of heart sound artifact from respiratory signals using singular spectrum based advanced line enhancer,” International Journal on Advanced Computer Theory and Engineering, vol. 2, no. 4, pp. 107–112, 2013.

[11] R. A. Thuraisingham, ”Preprocessing RR interval time series for heart rate variability analysis and estimates of standard deviation of RR intervals,” Computer Methods and Programs in Biomedicine, vol. 83, no. 1, pp. 78–82, 2006.

[12] M. R. Allen and L. A. Smith, “Investigating the origins and significance of low-frequency modes of climate variability,” Geophysical Research Letters, vol. 21, no. 10, pp. 883–886, 1994.

[13] A. L. Goldberger, L. A. Amaral, L. Glass et al., ”PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, pp. E215–E220, 2000.

[14] http://www.statlab.uni-heidelberg.de/people/eichler/timeseries/ts5.pdf.

[15] H. Hassani, R. Mahmoudvand, M. Zokaei, and M. Ghodsi, “On the separability between signal and noise in singular spectrum analysis,” Fluctuation and Noise Letters, vol. 11, Article ID 1250014, 25 pages, 2012.

[16] N. Golyandina, ”On the choice of parameters in singular spectrum analysis and related sub space-based methods,” Statistics and Its Interface, vol. 3, no. 3, pp. 259–279, 2010.

[17] A. Musialik-Lydko, B. Sredniawa, and S. Paszyk, ”Heart rate variability in heart failure,” Polish Heart Journal, vol. 58, no. 1, pp. 10–16, 2003.

[18] M. Costa, A. L. Goldberger, and C.-K. Peng, ”Multiscale entropy analysis of complex physiologic time series,” Physical Review Letters, vol. 89, no. 6, Article ID 068102, 4 pages, 2002.
