Detection of Nonlinearity and Stochastic Nature in Time Series by Delay Vector Variance Method

Imtiaz Ahmed

Abstract -- This paper investigates the suitability of Delay Vector Variance (DVV) algorithm in determining the presence of non-linearity and stochastic nature of time series both in the presence and absence of chaos. A differential entropy based method is used to find the optimum embedding dimension and time lag which are needed to represent the time series in phase space. Results obtained from the simulation indicate that the DVV method gives quantitative and easy to interpret output to characterize the underlying times series in terms of non-linear and stochastic nature.

Index Term -- Delay Vector Variance (DVV), Entropy Ratio (ER), Nonlinearity, Phase Space.

I. INTRODUCTION

Time series becomes chaotic due to the presence of determinism, non-linearity and strange attractor in it. So chaos implies non-linearity but not vice-versa. Since many non-linearity analysis techniques rest upon chaos theory, the properties such as determinism and presence of strange attractor have often been confounded with the non-linear behavior. So the presence of strange attractor would lead to the conclusion that the time series is non-linear while it is not necessarily so. As a result determinism and non-linearity have been confounded in time series signal analysis. Methods introduced by Kaplan[1] and by Kennel et al.[2] both rest upon the examination of the predictability of a time series. Furthermore, Kaplan’s ‘δ-ε’ method has been used in combination with the surrogate data [3] strategy for examining the linear or non-linear nature of time series. Kaplan’s method has short comings. It’s main shortcoming is in the large spread of results for surrogate data which tends to decrease the significance of results found with delta-epsilon. There is also Surrogate method [4] of testing non-linearity but there is a chance of false rejection of the null hypothesis due to stringent definition of linearity of the time series itself. Another classic method to detect non-linearity is ‘Deterministic versus Stochastic’ (DVS) [5] plots. But DVS method does not allow for a quantitative analysis. Another approach to non-linearity detection is Correlation Exponents (COR) [6]. It examines the local structure of a strange attractor. But linear time series which do exhibit strange attractor or geometric structure in phase space can be erroneously judged non-linear. To this cause a unified approach to analyze the predictability and the degree of non-linearity in time series signal was needed. This paper investigates such a unified approach to characterize time series namely the ‘Delay Vector Variance’ (DVV) method [7]. This method requires the underlined time series to be represented in phase space with optimum embedding dimension and time lag. A Differential Entropy based method [8] has been used here to find the optimum embedding dimension and time lag for time series representation in phase space.

II. BACKGROUND

A time series is a collection of observations made sequentially in time. There are four fundamental properties that describe the nature of the time series: linear, non-linear, deterministic and stochastic properties. Time series should be characterized in terms of these properties so that modeling, prediction and estimation can be done to different physical phenomena. By definition, a time series is considered linear if it can be generated by passing Gaussian white noise through a linear time invariant system. A time series is called deterministic if it can be uniquely defined by an explicit mathematical expression, a table of data, or a well defined rule. All past, present and future values of this type of time series are known precisely without any uncertainty. A time series is considered stochastic if its analysis and description involve uncertainty and unpredictability and need statistical techniques instead of explicit formulas.

III. DELAY VECTOR VARIANCE METHOD

Time Delay Embedding: It is possible to represent a time series in so called ‘phase space’ by the method of time delay embedding. When time delay is embedded into a time series it can be represented by a set of delay vectors (DVs) of a given dimension. If m is the dimension of the delay vectors then it can be expressed as x(k) = [x_{k-m},...,x_{k-1},T], where τ is the time lag. Now for every DV x(k), there is a corresponding target, namely the next sample x_{k+1}.

Characterization of Time series: A set λ_k is generated by grouping those DVs that are within a certain Euclidean
distance to DV x(k). This Euclidean distance will be varied in a manner standardized with respect to the distribution of pairwise distances between DVs. Now for a given embedding dimension m, a measure of unpredictability \( \sigma^2 \) is computed over all sets of \( \lambda_k \). The variation of the standardized distance enables the complete range of pairwise distances to be examined.

Now the algorithm for Delay Vector Variance (DVV) Method is summarized below:

For a given embedding dimension m, delay vectors (DV) \( x(k) = \{x_{k-n}, \ldots, x_k\}^T \) and corresponding targets \( x_k \), are generated:

The mean \( \mu_d \) and the standard deviation \( \sigma_d \) are computed over all pair wise Euclidean distances between DVs given by

\[
d(i, j) = \|x(i) - x(j)\| \quad (i \neq j).
\]

The sets \( \lambda_k(r_d) \) are generated such as that \( \lambda_k(r_d) = \{x(i) : \|x(k) - x(i)\| \leq r_d\} \), i.e., sets which consist of all DVs that lie closer to \( x(k) \) than a certain distance \( r_d \), taken from the interval \([\min(0, \mu_d - n_d \sigma_d), \mu_d + n_d \sigma_d]\) where \( n_d \) is a parameter controlling the span over which to perform DVV analysis.

For every set \( \lambda_k(r_d) \), the variance of the corresponding targets \( \sigma^2_k(r_d) \) is computed. The average over all sets \( \lambda_k(r_d) \), normalized by the variance of the time series signal, \( \sigma^2 \), yields the ‘target variance’

\[
\sigma^2_k(r_d) = \frac{\sum_{k=1}^{N} \sigma^2_k(r_d)}{N}
\]

where \( N = \text{Total number of sets } \lambda_k(r_d) \)

Variance measurement is assumed to be valid, if the set \( \lambda_k(r_d) \) contains at least 30 DVs, since having too few points for computing a sample variance yields unreliable estimates of the true population variance. This is considered as a minimum set size \( N_o \).

As a result of the standardization of the distance axis \( r_d \) is replaced by \( \frac{r_d - \mu_d}{\sigma_d} \) which will have zero mean and unit variance. Then the ‘DVV plots’ which are target variance \( \sigma^2_k(r_d) \) as a function of the standardized distance are quite easy to interpret. The minimum target variance, \( \sigma^2_{\text{min}} = \min_{r_d} [\sigma^2_k(r_d)] \) is a measure of noise present in the time series. The amount of noise is the prevalence of the stochastic component. The presence of strong deterministic component will lead to small target variance for small spans. At the extreme right, the DVV plots smoothly converges to unity, since for maximum spans, all DVs belong to the same universal set, and the variance of the targets is equal to the variance of the time series.

In order to generate surrogate time series, Iterative Amplitude Adjusted Fourier Transform (IAAFT) [9] has been used. With the help of the original and the surrogate time series DVV scatter diagram can be produced, where the horizontal axis corresponds to the DVV plot of the original time series and the vertical to that of the surrogate time series. DVV plots can be interpreted as follows.

If the surrogate time series signal yields DVV plots similar to that of the original time series, the ‘DVV scatter diagram’ coincides with the bisector line and original time series is judged to be linear. If the surrogate time series yield DVV plots not similar to that of the original time series, the curve will deviate from the bisector line and original time series is judged to be non-linear. Thus the deviation from the bisector line is an indication of non-linearity, and can be quantified as root mean square error (RMSE) between the \( \sigma^2 \)s of the original time series and the \( \sigma^2 \)s averaged over the DVV plots of the surrogate time series:

\[
t^\text{DVV} = \sqrt{\frac{\sum_{i=1}^{N^s} \sigma^2_{d,i}^2}{N^s}}
\]

Where \( \sigma^2_{d,i} \) is the target variance at span \( r_d \) for the i-th surrogate and the average is taken over all span of \( r_d \) that is valid in all surrogate and DVV plots.

IV. THE ENTROPY RATIO (ER) METHOD

The amount of disorder in time series can be measured by the Kozachenko-Leonenko (K-L) [10] estimate of the differential entropy

\[
H \approx \sum_{j=1}^{N} \ln(N) \rho_j + \ln 2 + C_E
\]

where \( N \) is the number of samples in the time series, \( \rho_j \) is the Euclidean distance of the j-th delay vector to its nearest neighbor, and \( C_E = 0.5772 \) is the Euler constant. For a given embedding dimension \( m \), and time lag \( \tau \), let \( H(x, m, \tau) \) denote the differential entropies estimated for time delay embedded versions of a time series, \( x \). The set of optimal parameters \( \{m_{\text{opt}}, \tau_{\text{opt}}\} \), yields a phase space representation which best reflects the dynamics of the underlying signal production system. Therefore, it is expected that this representation has minimal differential entropy (minimal disorder). Thus, optimization of the differential entropy (Eq. 4) for \( m \) and \( \tau \) is achieved and the minimum of \( H(x, m, \tau) \), yielding the optimal set of embedding parameters \( \{m_{\text{opt}}, \tau_{\text{opt}}\} \) is obtained. The K-L estimates for the time delay embedded versions of the original time series \( H(x, m, \tau) \), and its surrogates \( H(x_{\tau,i}, m, \tau) \) are computed using Eq. 4 for increasing \( m \) and \( \tau \) (index \( i \) refers to the i-th surrogate). To determine the optimal embedding parameters, the ratio

\[
I(m, \tau) = \frac{H(x, m, \tau)}{H(x_{\tau,i}, m, \tau)}
\]

needs to be minimized, where \( \langle \cdot \rangle_i \) denotes the average over \( i \). To penalise for higher embedding dimensions, the minimum description length (MDL) method is superimposed, yielding the “entropy ratio” (ER):

\[
R_{\text{ent}}(m, \tau) = I(m, \tau)(1 + \min_{i} N_i)
\]

where \( N_i \) is the number of delay vectors, which is kept constant for all values of \( m \) and \( \tau \) under consideration.

V. CONSTRUCTION OF TIME SERIES

To test the effectiveness of the Delay Vector Variance Method, a number of time series [11,12] have been
generated which have different degrees of nonlinearity and stochastic nature.

Model 1: Chaotic Mackey-Glass time series which is extensively used in non-linear dynamical system modeling is defined by:

\[
\frac{dx}{dn} = \frac{0.2x_{n-\tau}}{1 + x_{n-\tau}^{10}} - 0.1x_n
\]  

(7)

We use \(x_0=0.2\) and \(\tau=17\) to generate the time series.

Model 2: Chaotic Lorenz time series which is widely used in continuous dynamical system is considered. This time series is defined by

\[
\begin{align*}
\frac{dx}{dt} &= \delta(y-x) \\
\frac{dy}{dt} &= rx - y - xz \\
\frac{dz}{dt} &= -bx + xy 
\end{align*}
\]

(8)

where \(\delta = 10\), \(r = 28\) and \(b = \frac{8}{3}\). The sampling frequency is taken to be 100 Hz (oversampled) to generate the time series.

Model 3: Hénon map can be written in terms of a single variable with two time delays:

\[
x_{n+1} = 1 + ax_n^2 + bx_{n-1}
\]

(9)

where \(a=1.4\) and \(b=0.3\). This ensures the Hénon map to be chaotic. This map is widely used in weather forecasting.

Model 4: A time is generated by a fully deterministic chaotic Feigenbaum recursion of the form

\[
X_n=3.57X_{n-1}(1-X_{n-1})
\]

(10)

where the initial condition is, \(X_0=0.7\). The stochastic nature is absent here since the model is noise free. Also this model exhibits third order non-linear dependence.

Model 5: A time series is generated by an Autoregressive Conditional Heteroscedastic (ARCH) process of the form

\[
X_n = (1+0.5X_{n-1}^2)^{1/2} \eta_n
\]

(11)

with the value of the initial observation set to \(X_0 = 0\). The sequence \(\{ \eta_n \}\) are sampled from standard normal distribution. This model is a statistical forecasting model for volatility. It considers the variance of the current error term to be a function of the variances of the previous time periods error terms. It is employed commonly in modeling financial time series that exhibits time varying volatility clustering, i.e. periods of swing followed by periods of relative calm.

Model 6: A time series is generated by a nonlinear moving average process (NLMA) of the form

\[
X_n = \eta_n + 0.8 \eta_{n-1} \eta_{n-2}
\]

(12)

where the sequence \(\{ \eta_n \}\) are sampled from standard normal distribution.

Model 7: A time series is generated by a linear stochastic autoregressive moving average process (ARMA) of the form:

\[
X_n = 0.8X_{n-1} + 0.15X_{n-2} + \eta_n + 0.3 \eta_{n-1}
\]

(13)

with \(X_0=1\), \(X_1=0.7\) and the sequence \(\{ \eta_n \}\) is sampled from standard normal distribution. This model contains both moving average and autoregressive terms.

VI. Simulations

In this section, the Differential entropy based method is used to find out the optimum embedding dimension \(m_{opt}\) and the time lag \(\tau_{opt}\) in order to represent the time series generated by models 1 to 7 in phase space. The Entropy-Ratio plot (ER-plot) of the different time series mentioned above is shown in Fig. 1 to Fig. 7. These optimum values will be used in DVV method for determining the non-linear and stochastic nature in the time series.
Table I, gives the \{ m_{opt}, \tau_{opt} \} for all the time series. Now, using these parameters we use DVV method to find out the presence of non-linearity and stochastic nature.

Simulation parameters for DVV method are given as: Minimal Size of Set, \( \lambda_k (r_d), N_0 = 30 \), Number of surrogates, \( N_s = 19 \), Maximal Span, \( n_d = 3 \), Number of Reference, DVs = 500, Embedding Dimension, \( m \) and Time Delay, \( \tau \) are chosen according to Table I, Sample point numbers\( = 2000 \).
VII. DISCUSSION

The first four time series namely Mackey-Glass, Lorenz, Hénon and Feigenbaum are chaotic. It ensures the presence of strange attractor in their phase space representation, strong deterministic nature and the non-linearity. In the DVV plots of these time series in Fig. 8 to Fig. 11, very small value for the target variance $\sigma^2_{min}$ (0.09469, 0.1785, 0.005027, 0.0005037 successively) for original series indicates the presence of strong deterministic component since $\sigma^2_{min}$ is a measure of noise present in the time series and the amount of noise is the prevalence of the stochastic component. Specially for chaotic Feigenbaum Recursion target variance $\sigma^2_{min}$ is 0.0005037 ≈ 0. This indicates fully deterministic behavior. From the DVV scatter diagram of...
these time series which are shown in Fig.15 to Fig.18, we can conclude that the DVV method has successfully detected the presence of non-linearity as the curves in these figures deviate significantly from the bisector line. This is because the $\sigma^2$ for surrogates are different than that of original time series. For the rest of the three time series namely ARCH, NLMA and ARMA do not have chaotic behavior but non-linearity and stochastic component are present in ARCH and NLMA. DVV method can still successfully indicates the presence of strong stochastic nature in these time series by giving a large value of target variance $\sigma^2_{\min} = 0.6493$. Also, in Fig.19 and Fig.20, the DVV scatter diagram significantly deviates from the bisector line which indicate the non-linear nature in NLMA and ARCH. Finally the time series generated by ARMA process is stochastic but linear. This is clearly shown in Fig.14 where the $\sigma^2_{\min}$ is 0.1041 which indicates the presence stochastic nature and in Fig.21 linearity is judged as the DVV scatter diagram coincides with the bisector line. So Delay Vector Variance method can successfully determine the presence of nonlinear and stochastic nature in a time series regardless of the fact that whether or not chaos is present. This is possible because DVV method depends on the local predictability of the time series in phase space which can be observed by the target variance of delay vectors (DVs).

REFERENCES