Combined use of correlation dimension and entropy as discriminating measures for time series analysis

K.P. Harikrishnan\textsuperscript{a,}\textsuperscript{,}, R. Misra\textsuperscript{b}, G. Ambika\textsuperscript{c}

\textsuperscript{a}Department of Physics, The Cochin College, Cochin 682002, India
\textsuperscript{b}Inter-University Center for Astronomy and Astrophysics, Pune 411007, India
\textsuperscript{c}Indian Institute of Science Education and Research, Pune 411021, India

\begin{abstract}
We show that the combined use of correlation dimension ($D_2$) and correlation entropy ($K_2$) as discriminating measures can extract a more accurate information regarding the different types of noise present in a time series data. For this, we make use of an algorithmic approach for computing $D_2$ and $K_2$ proposed by us recently [Harikrishnan KP, Misra R, Ambika G, Kembhavi AK. Physica D 2006;215:137; Harikrishnan KP, Ambika G, Misra R. Mod Phys Lett B 2007;21:129; Harikrishnan KP, Misra R, Ambika G. Pramana – J Phys, in press], which is a modification of the standard Grassberger–Procaccia scheme. While the presence of white noise can be easily identified by computing $D_2$ of data and surrogates, $K_2$ is a better discriminating measure to detect colored noise in the data. Analysis of time series from a real world system involving both white and colored noise is presented as evidence. To our knowledge, this is the first time that such a combined analysis is undertaken on a real world data.
\end{abstract}

\section{1. Introduction}

The first major step to investigate the existence of nontrivial structures in a time series data is by means of hypothesis testing using surrogate data \cite{1,2}. To test the null hypothesis, a method sensitive to nonlinearity is applied to the original time series and to an ensemble of surrogate data. The null hypothesis is rejected if a discriminating statistic derived from the method statistically discriminates the original from the surrogates. Traditionally, the correlation dimension $D_2$ is considered as the most useful measure for making such distinctions. Though there are other important measures such as, for example, Lyapunov exponent and correlation entropy, they are not often used in practice for hypothesis testing as they require comparatively more number of data points and are also computationally more complex. A comparative study on the discriminating powers of different measures for nonlinearity in a time series is given in Schreiber and Schmitz \cite{3}.

One apparent difficulty for using $D_2$ as a discriminating measure is when the time series involves colored noise. Since colored noise is a correlated random process, it gives a well saturated value of $D_2$, just like a chaotic data \cite{4}. In this context, the correlation entropy $K_2$ assumes more significance as a discriminating measure, as suggested by many authors \cite{5–7}. While $D_2$ is a static measure characterising the geometry of the attractor, $K_2$ is a dynamic measure representing the rate at which information needs to be created as the chaotic system evolves in time \cite{8}. Thus, in contrast to chaotic systems (where $K_2$ saturates as the embedding dimension $M$ increases), $K_2 \rightarrow 0$ as $M \rightarrow \infty$ for colored noise \cite{9}. Though the importance of $K_2$ has been realised for analysing time series involving colored noise, this has not been explicitly shown through surrogate
analysis for any practical data. One reason for this is the difficulties associated with the conventional method for computing $K_2$.

The standard method for computing both $D_2$ and $K_2$ is the Grassberger–Proccacia [GP] algorithm [10,11], even though a few other methods have also been proposed in the literature to compute $D_2$ [12,13] and $K_2$ [14] for specific data sets. The technique uses the scalar time series to reconstruct the dynamics in an embedding space of dimension $M$ using delay coordinates scanned at a suitable time delay $\tau$. The procedure also involves a visual identification of the scaling region in the correlation sum, which becomes more difficult for practical data involving noise.

To conduct the surrogate analysis, the $D_2$ of the data and the surrogates are determined by identifying the scaling region subjectively. One basic assumption for the success of surrogate analysis is that the same conditions apply for the data and the surrogates while computing the test statistic. But the subjectivity in the computation of $D_2$ and $K_2$ does not allow one to be certain that exactly the same criteria have been applied to analyse both the real and the surrogate data, which is crucial in testing the null hypothesis. Even though the rejection is statistical, based on the significance level, the effectiveness of the method may be reduced when noise is present.

To overcome these difficulties, we have recently proposed and implemented an algorithmic approach for computing $D_2$ [15] as well as $K_2$ [16]. In our scheme, the scaling region is determined algorithmically based on some criteria which remain the same for the data and the surrogates. It should be emphasized that we have only modified the original GP algorithm with a nonsubjective computation of the scaling region using some criteria imposed in the algorithm. Here we show that this enables a combined and more efficient analysis of the time series with both $D_2$ and $K_2$ as discriminating measures. Results are explicitly presented for real world data involving white and colored noise.

Our paper is organised as follows: In Section 2, the details regarding the computation of $D_2$ and $K_2$ are given. The generation of surrogate data and the numerical results are presented in Section 3 and the paper is concluded in Section 4.

2. Computation of $D_2$ and $K_2$

In this section, we present the essential details for computing $D_2$ and $K_2$ required for this analysis. A more complete discussion with application to more number of chaotic systems is presented elsewhere for both $D_2$ [15] and $K_2$ [16].

The delay embedding technique uses an embedding space of dimension $M$ with delay vectors constructed by splitting a discretely sampled scalar time series $s(t)$ with delay time $\tau$ as

$$\mathbf{x}_i = [s(t_i), s(t_i + \tau), \ldots, s(t_i + (M - 1)\tau)].$$

(1)

The correlation sum is the relative number of points within a distance $R$ from a particular $(i$th) data point,

$$p_i(R) = \lim_{N_i \to \infty} \frac{1}{N_i} \sum_{j \neq i} H(R - |\mathbf{x}_i - \mathbf{x}_j|).$$

(2)

where $N_i$ is the total number of reconstructed vectors and $H$ is the Heaviside step function. Averaging this quantity over $N_i$ randomly selected $\mathbf{x}_i$ or centers gives the correlation function

$$C_M(R) = \frac{1}{N_i} \sum_{i} p_i(R).$$

(3)

The correlation dimension $D_2(M)$ is then defined to be

$$D_2 = \lim_{R \to 0} d(\log C_M(R))/d(\log(R)).$$

(4)

which is the scaling index of the variation of $C_M(R)$ with $R$ as $R \to 0$.

Also, as $M$ increases, one expects $C_M(R)$ to decrease for a fixed value of $R$. This is because, the computation of $C_M(R)$ involves how many trajectory points in the embedding space stay within the distance $R$ of each other. The correlation entropy $K_2$ is defined by the relation

$$C_M(R) \approx e^{-MK_2\Delta t},$$

(5)

where $\Delta t$ is the time step between successive values in the time series. From this, a formal expression for $K_2$ can be written as

$$K_2\Delta t = \lim_{R \to 0} \lim_{M \to \infty} \lim_{N_i \to \infty} \log(C_M(R)/C_{M+1}(R)).$$

(6)

To compute both $D_2$ and $K_2$, the conventional method is to identify a scaling region in the correlation sum by visual inspection. However, such an exercise is subjective and may not be ideal for surrogate analysis where, the same conditions are to be applied on the data and the surrogates for proper comparison.

Hence we adopt a modification in the computational procedure where, the scaling region is fixed algorithmically. For this, a maximum value of $R$ ($R_{max}$) and a minimum value of $R$ ($R_{min}$) are computed for each $M$ using some criteria [15,16] based on the algorithm itself and the region between them is taken as the scaling region. For a fixed $M$ value, $D_2$ and $K_2$ are calculated for different values of $R$ in the scaling region using Eqs. (4) and (6), respectively, and the average is calculated. The error in $D_2$ and $K_2$ are also estimated as the mean standard deviation over the average value.
3. Results of surrogate analysis

The rationale behind surrogate analysis is to formulate a null hypothesis that the data have been created by a stationary linear stochastic process, and then to attempt to reject this by comparing a suitable measure for the data with appropriate realisations of surrogate data. The rejection of null hypothesis is statistical based on a significance level $S$ given by

$$S = \frac{|f - \langle f \rangle_{\text{surr}}|}{\sigma_{\text{surr}}},$$

where $f$ is a suitable measure for the data, $\langle f \rangle_{\text{surr}}$ denotes the mean of the distribution of $f$ for surrogates and $\sigma_{\text{surr}}$ its standard deviation. But here we use a quantitative measure that is slightly different as discussed below.

The method for the generation of surrogate data was originally proposed by Theiler and coworkers [1], with the Amplitude Adjusted Fourier Transform (AAFT) algorithm. Later another algorithm was proposed by Schreiber and Schmitz [17,18] which is similar, but makes use of an iterative scheme in order to achieve arbitrary close approximation to the autocorrelation and amplitude distribution. It is referred to as the iterative algorithm (IAAFT) and has been shown to be more consistent in testing null hypothesis [19] for a wide class of stochastic process. Hence in this work, we apply this scheme to generate surrogate data sets using the details given in [20]. All the analysis are done with 30,000 long data points with 10 surrogates for each data.

We first test our numerical scheme using time series from a standard low dimensional chaotic attractor, namely the Rossler attractor, at parameter values $a = b = 0.2$ and $c = 7.8$. Data consist of 30,000 points generated with the time step $\Delta t = 0.1$. Ten surrogates are also generated for the data using the IAAFT scheme. The result of computation of $D_2$ and $K_2$ for the data and surrogates are shown in Figs. 1 and 2, respectively. In all the figures, the values of the surrogates are shown by dashed lines without error bar for clarity.

The real world data are always contaminated by noise and the question that arises naturally is how much amount of noise can suppress the nonlinear component that may be present in a time series. The power in a noise process varies in general as $1/f^a$, where the value of $a$ determines the type of noise. For white noise, $a = 0$ and for colored noise, it varies from 1.0 to 2.0. For the analysis here, we choose colored noise with $a = 2.0$, which is called the red noise. Four data sets are generated from the Rossler data by adding 20% and 50% of white noise and the same percentages of red noise. The surrogates are also generated for each data. Results of applying our numerical scheme to each of these data and their surrogates are shown in Figs. 3 and 4.

By comparing the two figures, it becomes clear that the behavior of $D_2$ and $K_2$ are complimentary to each other with respect to white noise and colored noise contamination. As the percentage of white noise increases, $K_2$ saturates at higher val-

![Fig. 1. Saturated value of $D_2$ as a function of embedding dimension $M$ for the time series from Rossler attractor (circles with error bar) and its 10 surrogates (dashed lines). The parameter values used for computation are $a = b = 0.2$, $c = 7.8$ and the number of data points used are 30,000.](image-url)
Fig. 2. Saturated value of $K_2$ (circles) as a function of $M$ for the Rossler attractor data used in the previous figure and 10 surrogates.

Fig. 3. The upper panel shows $D_2$ as a function of $M$ for the data obtained by adding (a) 20% white noise and (b) 50% white noise to the Rossler attractor data and 10 surrogates for both. The lower panel shows the same results, but for data obtained by adding (c) 20% red noise and (d) 50% red noise.
ues while $D_{sat}^2 \rightarrow 1$ and becomes indistinguishable from the surrogates. The opposite happens in the case of colored noise contamination. With increase in colored noise, $D^2$ always saturate; but $K_{sat}^2 \rightarrow 0$ along with the surrogates. Thus, while white noise contamination can be easily identified through $D^2$ analysis, the presence of colored noise can be inferred better by computing $K^2$.

The above conclusions appear qualitative and a quantification is attempted using a measure we call the normalised mean sigma deviation (nmsd). For $D^2$, this is defined as

$$nmsd^2 = \frac{1}{M_{max} - 1} \sum_{M=2}^{M_{max}} \frac{D^2(M) - \langle D^2_{sur} \rangle(M)}{\sigma^2_{sur}^2(M)}^2$$

(8)

with a similar expression for $K^2$. Here, $\langle D^2_{sur} \rangle(M)$ is the average $D^2$ for the surrogate data for each $M$, $M_{max}$ is the maximum embedding dimension for which the analysis is undertaken (which is 8 here) and $\sigma^2_{sur}^2(M)$ is the standard deviation of $D^2_{sur}(M)$. For the Rossler attractor, nmsd is equal to 24.6 and 32.2 for $D^2$ and $K^2$, respectively, as computed from Figs. 1 and 2. With $D^2$ as discriminating measure (Fig. 3), the values for 20% and 50% white noise contamination are 9.6 and 1.8, while for red noise, the respective values are 12.4 and 3.6. Similarly, with $K^2$ as discriminating measure (Fig. 4), the values are 8.4 and 4.3 (for white noise) and 5.6 and 2.2 (for red noise). Thus a small value for nmsd (say, <3.0) with respect to either $D^2$ or $K^2$ indicates a high percentage of noise contamination in the data.

Finally, we show the effectiveness of a combined analysis with $D^2$ and $K^2$ as discriminating measures to extract information regarding the presence of possible nonlinear structures in real world data. We analyse the light curve from an astrophysical X-ray source, namely, GRS1915+105. The temporal behavior of this black hole system has been classified into 12 different states [21], and many of these states are expected to contain different percentages of white and colored noise. Here we choose data from four representative states, namely, $\theta$, $\beta$, $\kappa$ and $\gamma$ and generate ten surrogates for each data. The data from each state contain continuous data streams with 30,000 data points. More details regarding the data can be found elsewhere [22]. The results of a combined analysis of the data with $D^2$ and $K^2$ as discriminating measures are shown in Figs. 5 and 6, respectively. In Fig. 5, with $D^2$ as discriminating measure, all states except the $\gamma$ state show deviation from stochastic behavior. But when $K^2$ is used as discriminating measure (Fig. 6), the $\kappa$ state clearly shows colored noise contamination, which is not evident from Fig. 5 with $D^2$ as discriminating measure. This indicates the importance of a combined analysis with more than one discriminating measure to get a more accurate information regarding the nature of the time series.

**Fig. 4.** Results of computation of $K^2$ as a function of $M$ for the four data sets and their surrogates analysed in the previous figure and shown in the same order for white and red noise contamination. Note that the red noise contamination can be more easily identified by computing $K^2$ of the data and the surrogates.
Fig. 5. Results of surrogate analysis with $D^2$ as the discriminating measure for the four temporal states – $b$, $k$, $\theta$, $\gamma$ – of the black hole system GRS1915+105. The null hypothesis can be rejected for the first three states.

Fig. 6. Results of surrogate analysis of the four black hole states with $K^2$ as the discriminating measure. Note that variation of $K^2$ for the $\kappa$ state indicates colored noise contamination which is not evident from the previous figure.
We have also computed the \( nmsd \) for the four states. It is found that for the states \( \beta \) and \( \delta \), \( nmsd > 5.0 \) with respect to both \( D_2 \) and \( K_2 \) while for \( \gamma \) state, \( nmsd < 3.0 \) in both cases. But for \( \kappa \) state, \( nmsd \) is 4.6 for \( D_2 \) and 3.3 for \( K_2 \). This confirms colored noise contamination for the state.

4. Conclusion

The presence of nontrivial structures in a time series data is often inferred through surrogate analysis with \( D_2 \) as discriminating measure. Here we show explicitly that the combined use of \( D_2 \) and \( K_2 \) as discriminating measure is more effective, especially when the data is expected to contain colored noise. To implement the procedure in a more efficient manner, an algorithmic approach for computing \( D_2 \) and \( K_2 \) is prescribed, which ensures that the same conditions are maintained in the algorithm for both data and surrogates. A quantification for discrimination between data and surrogates is attempted using \( nmsd \) and results are explicitly shown for practical data involving white and colored noise.

It should be emphasized that we are not proposing any new method for surrogate analysis, nor do we claim to establish low dimensional chaos by this method. We have only shown that by automating the conventional method of computing \( D_2 \) and \( K_2 \), surrogate analysis can be done more efficiently with both as discriminating measures. This, in turn, enables a more complete analysis of the data using the two complementary measures of low dimensional chaos.

Identifying nontrivial behavior in real world systems is of great practical importance. But the fact that null hypothesis can be rejected is by no means an assurance for the presence of low dimensional chaos in the underlying process, as this is only one possibility to violate null hypothesis. In principle, it requires a succession of tests using different types of surrogates and other important measures from nonlinear dynamics before making such conclusions. In this sense, our method do serve as an effective and reliable guide to indicate whether such refined analysis is really warranted on the data. In particular, it may be used along with other important quantifiers of chaos for getting more efficient and precise results.

Acknowledgement

KPH acknowledges the computing facilities in IUCAA, Pune.

References