PRAMANA
— journal of
physics

© Indian Academy of Sciences

Vol. 72, No. 2
February 2009
pp. 325–333

Efficient use of correlation entropy for analysing time series data

K P HARIKRISHNAN[1,*], R MISRA[2] and G AMBIKA[3]

[1]Department of Physics, The Cochin College, Cochin 682 002, India

[2]Inter-University Centre for Astronomy and Astrophysics, Ganeshkhind, Pune 411 007, India

[3]Indian Institute of Science Education and Research, Pune 411 021, India

[*]Corresponding author. E-mail: kp_hk2002@yahoo.co.in

MS received 8 January 2008; revised 9 October 2008; accepted 14 October 2008

Abstract. The correlation dimension D_2 and correlation entropy K_2 are both important quantifiers in nonlinear time series analysis. However, use of D_2 has been more common compared to K_2 as a discriminating measure. One reason for this is that D_2 is a static measure and can be easily evaluated from a time series. However, in many cases, especially those involving coloured noise, K_2 is regarded as a more useful measure. Here we present an efficient algorithmic scheme to compute K_2 directly from a time series data and show that K_2 can be used as a more effective measure compared to D_2 for analysing practical time series involving coloured noise.

Keywords. Time series analysis; correlation entropy; coloured noise.

PACS Nos 05.45.Ac; 05.45.Tp; 05.40.Ca

1. Introduction

Nonlinear time series analysis is the most effective link between chaos theory and the real world. It has now been realized that detecting nontrivial structures in experimental time series requires a succession of tests using various measures. Though a number of important nonlinearity measures have been identified to analyse time series data, the most important among them are the correlation dimension D_2 and the correlation entropy K_2. While the former is a static measure characterizing the structure of the chaotic attractor, the latter is a dynamic measure and represents the rate at which information needs to be created as the chaotic system evolves in time [1]. This is because, due to the sensitivity to initial conditions in chaotic systems, as the orbits evolve, initially insignificant bits in the specification of initial conditions eventually become significant as time tends to ∞. Hence K_2 is also closely related to the Lyapunov exponent [LE] [2], which measures the exponential rate of divergence of nearby trajectories in phase space.

Traditionally, K_2 has been much less popular compared to D_2 as a discriminating statistic in analysing time series in practice. This is because their values are generally much harder to determine [1] and requires much more number of data points for a reasonable estimate, compared to that of D_2. However, K_2 has a significant and more relevant status, especially in the context of coloured noise contamination, as indicated by many authors [3–5]. The standard method for both is the Grassberger–Proccacia (GP) algorithm [6,7], even though a few other methods have also been proposed in the literature to compute D_2 [8,9] and K_2 [10] 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 τ.

But a major difficulty in implementing this procedure is that, the scaling region in the correlation sum for the computation of D_2 and K_2 has to be identified subjectively, as discussed in detail in the next section. The problem becomes worse in the case of experimental time series due to various factors such as the limitation in the number of data resulting in edge effects, presence of noise etc. Though a number of improvements have been suggested, especially for the computation of D_2 [11–13], the problem of subjectivity of the scaling region still remained. To overcome this, we have recently proposed and implemented a modification in the computational scheme for GP algorithm to compute D_2, by identifying the scaling region algorithmically [14]. We have also shown that the method can be used for any arbitrary time series and is most suitable for hypothesis testing. A major purpose of this paper is to show that this scheme can be extended for the computation of K_2 as well and it provides a more effective use of K_2 for analysing time series involving coloured noise. The scheme is first verified using standard synthetic data sets and it is then applied to analyse experimental time series. Section 2 discusses the computational scheme in detail. The numerical results are presented in §3 and the conclusions are drawn in §4.

2. Computation of entropy

In this section, we present the essential details of our algorithmic scheme required for this analysis. A more complete discussion regarding the computation of D_2 is presented elsewhere [14]. The GP algorithm aims at creating an artificial space of dimension M with delay vectors constructed by splitting a discretely sampled scalar time series $s(t_i)$ with delay time τ as

$$\vec{x_i} = [s(t_i), s(t_i + \tau), ..., s(t_i + (M-1)\tau)]. \tag{1}$$

The correlation sum is the average number of points with the relative distance within R from a particular (ith) data point,

$$p_i(R) = \lim_{N_v \to \infty} \frac{1}{N_v} \sum_{j=1, j\neq i}^{N_v} H(R - |\vec{x_i} - \vec{x_j}|), \tag{2}$$

where N_v is the total number of reconstructed vectors and H is the Heaviside step function. Averaging this quantity over N_c randomly selected $\vec{x_i}$ or centres gives the correlation function

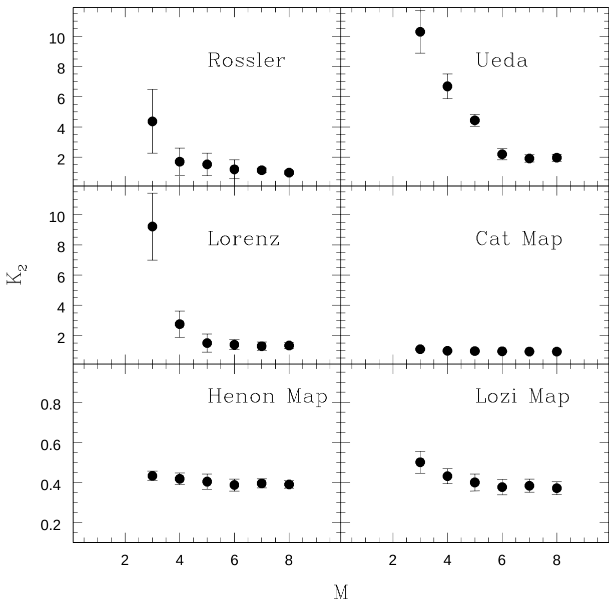

Figure 1. $K_2(M)$ values of six low-dimensional chaotic systems, each containing 30,000 data points. The saturated values are given in table 1.

$$C_M(R) = \frac{1}{N_c}\sum_i^{N_c} p_i(R). \tag{3}$$

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. As M increases, the lengths of the trajectory segments being compared are increased and K_2 measures the rate of change of this. Hence K_2 can be defined by the relation

$$C_M(R) \propto \mathrm{e}^{-MK_2\Delta t}, \tag{4}$$

where Δt is the time step between successive values in the time series. From above, a formal expression for K_2 can be written as

$$K_2\Delta t = \lim_{R\to 0}\lim_{M\to\infty}\lim_{N\to\infty}(-\log C_M(R)/M). \tag{5}$$

Alternately, K_2 can also be obtained as

$$K_2\Delta t \equiv \lim_{R\to 0}\lim_{M\to\infty}\lim_{N\to\infty}\log(C_M(R)/C_{M+1}(R)). \tag{6}$$

To compute K_2, one has to identify a linear part in the log $C_M(R)$ vs. log R plot for each M, called the scaling region, which is usually done by the visual inspection of the correlation sum. But in our computational scheme, this is avoided and instead, the scaling region is fixed algorithmically. For this, a maximum value of

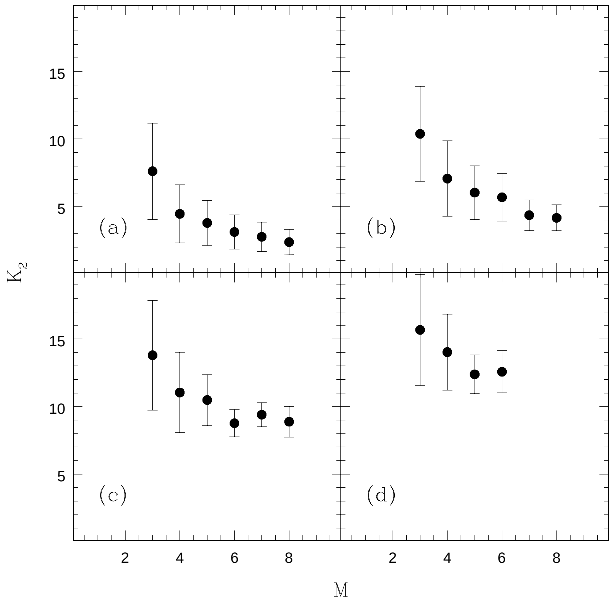

Figure 2. $K_2(M)$ values for four data sets obtained by adding (**a**) 10%, (**b**) 20%, (**c**) 50% and (**d**) 100% of white noise to the data from Rossler system. The value of K_2 and error bar increase with the increase in noise.

R, $R_{\rm max}$ and a minimum value of R, $R_{\rm min}$ are computed for each M using some criteria based on the algorithm itself and the region between them is taken as the scaling region. The criterion for $R_{\rm max}$ is that the number of available centres are at least $N_v/100$ and $R_{\rm min}$ is decided by the condition that on the average at least ten data points are considered per centre [14]. For a fixed M value, K_2 is calculated for different values of R in the scaling region using eq. (6) and the average is calculated. The error in $K_2(M)$ is also estimated as the mean standard deviation over this average value. There often exists a critical embedding dimension $M_{\rm cr}$ for which $R_{\rm min} \approx R_{\rm max}$, so that significant results can be obtained only for $M \leq M_{\rm cr}$. Hence the computations are done for each value of M starting from $M = 2$ to $M = M_{\rm cr}$.

3. Numerical results

To illustrate our scheme, we first analyse synthetic time series generated from six well-known low-dimensional chaotic systems. For all the analysis in this section, 30,000 data points are used. Figure 1 shows the $K_2(M)$ values computed for the six standard chaotic systems. In all cases, we get a well saturated value for $K_2(M)$. The saturated values $K_2^{\rm sat}$ are given in table 1 along with the corresponding standard values for comparison.

For the scheme to be useful in analysing real world data, it should effectively compute K_2 values of time series contaminated by noise. In order to show this, we

Table 1. $K_2^{\rm sat}$ values of six low-dimensional chaotic systems obtained using the scheme prescribed in this work along with the standard $K_2^{\rm sat}$ values given in literature.

System	Computed $K_2^{\rm sat}$	Standard $K_2^{\rm sat}$
Rossler attractor		
$(a = b = 0.2,\ c = 7.8)$	1.07 ± 0.13	1.04 ± 0.02
Lorenz attractor		
$(\sigma = 10, r = 28, b = 8/3)$	1.315 ± 0.18	1.327 ± 0.03
Ueda attractor		
$(k = 0.05, A = 7.5)$	1.92 ± 0.24	1.68 ± 0.13
Henon map		
$(a = 1.4, b = 0.3)$	0.39 ± 0.04	0.417 ± 0.06
Lozi map		
$(a = 1.7, b = 0.5)$	0.375 ± 0.032	0.39 ± 0.03
Cat map	0.965 ± 0.006	0.98 ± 0.02

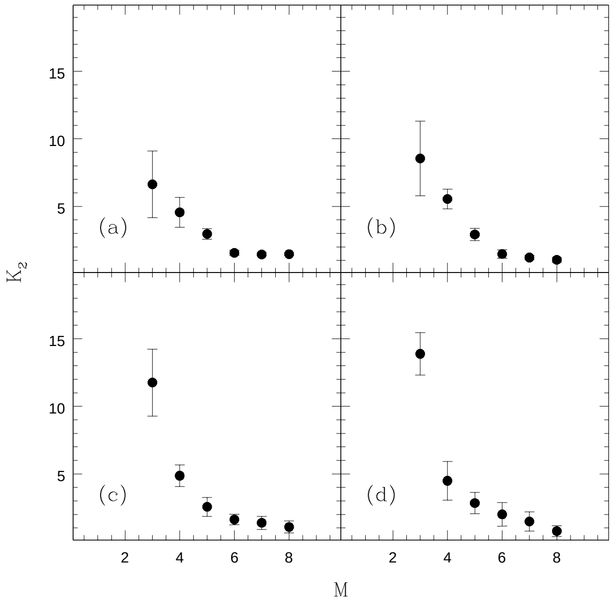

Figure 3. Same as figure 2, but with red noise added instead of white noise. Note that, in contrast to the previous figure, K_2 decreases with M as the level of red noise increases.

construct data sets by adding different percentages of white and coloured noise to the data from Rossler system. The power in a noise process varies in general as $1/f^{\alpha}$, where the value of α determines the type of noise. For white noise, $\alpha = 0$ and for coloured noise, it varies from 1.0 to 2.0. We choose coloured noise with $\alpha = 2.0$, which is called the red noise. For pure white noise, we expect $K_2 \rightarrow \infty$,

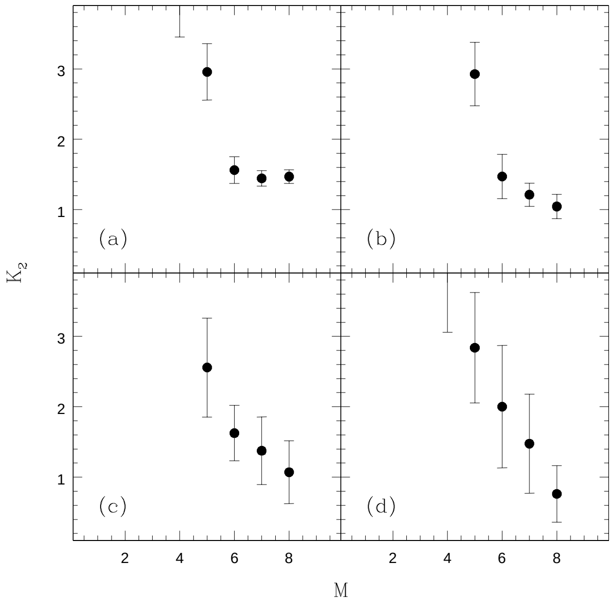

Figure 4. A magnified view of the region from $M = 5$ to 8 of figure 3, which clearly shows the dependence of K_2^{sat} on the level of red noise.

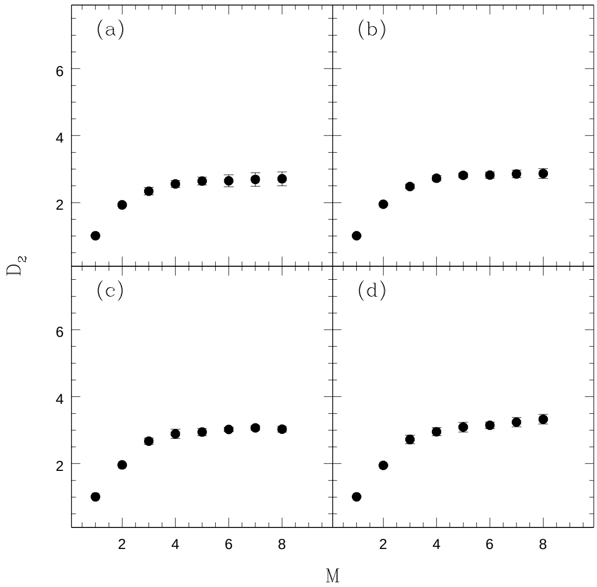

Figure 5. $D_2(M)$ values of the four data sets analysed in figure 3, involving different percentages of red noise. The saturated D_2 values are very close even with widely varying amounts of coloured noise.

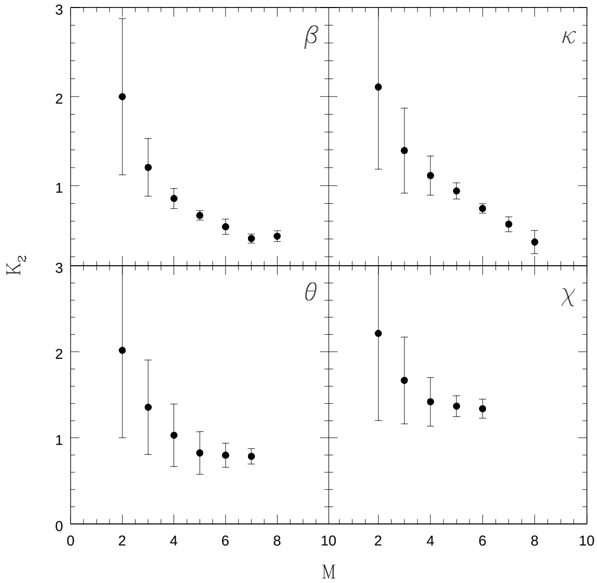

Figure 6. $K_2(M)$ values of four data sets corresponding to different temporal states of the black hole system GRS 1915+105. The values are computed per sec from 30,000 data points. Note that while θ and β indicate chaotic behaviour, κ clearly shows coloured noise contamination and χ is more like a white noise.

whereas, for coloured noise $K_2 \to 0$ as $M \to \infty$ [15], since it is a time-correlated random process.

Figure 2 shows the results of applying our scheme to four data sets with (a) 10%, (b) 20%, (c) 50% and (d) 100% of white noise added, whereas figure 3 shows the same results, but with red noise. As expected, $K_2^{\rm sat}$ increases as the level of white noise contamination increases, and $K_2^{\rm sat} \to 0$ with the increase in red noise. This is more clearly seen in figure 4 where a blow up of the region from $M = 5$ to 8 of figure 3 is shown. Moreover, the error bar in figure 2 also increases proportional to the white noise contamination. For a noise level of 100% in figure 2d, the scheme computes K_2 only up to a maximum M value of $M_{\rm cr} = 6$. It shows that beyond $M = 6$, there is no reasonable scaling region for the data.

In contrast, we show in figure 5, the D_2 values of the same data sets given in figure 3, having different percentages of coloured noise contamination. It is clear that, practically it is impossible to infer coloured noise contamination by computing D_2 alone.

Finally, we analyse a few data sets from an astrophysical X-ray source, the black hole system GRS 1915+105. All data sets contain continuous data streams with $N = 30{,}000$. Temporal behaviour of this black hole system has been classified into 12 different states and more details regarding this can be found elsewhere [16]. Here we choose data from four representative states, viz. θ, β, κ and χ, and the result of applying our scheme is shown in figure 6. Comparing this with the previous results on synthetic data added with noise, one finds that the states θ and β saturate much

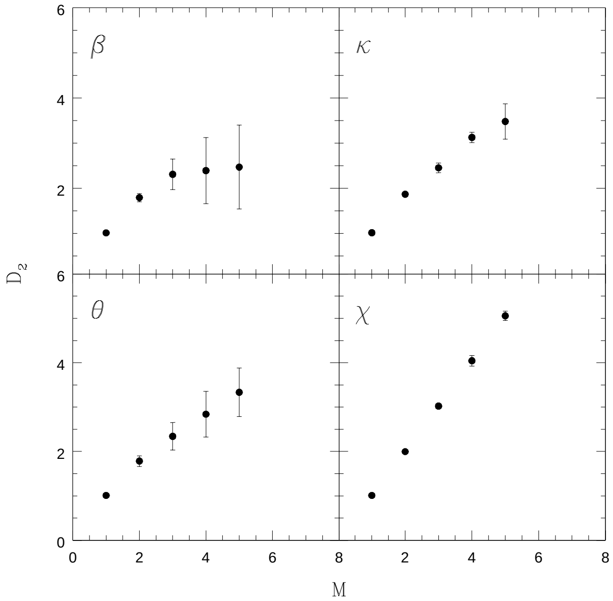

Figure 7. $D_2(M)$ values of the four states of the black hole system GRS 1915+105, whose $K_2(M)$ values are shown in figure 6.

like a chaotic system. The behaviour of κ indicates coloured noise contamination whereas χ can be identified as a white noise with much higher values of K_2 and computed only up to $M = 5$. Though these results agree with our previous analysis using D_2 [16], the effect of coloured noise on the κ state was not evident there, as can be explicitly seen from figure 7, where D_2 values of these states computed using our code are shown. This confirms the importance of K_2 as a measure in analysing time series involving coloured noise.

4. Conclusion

In this paper, we show that K_2 can be effectively used for the analysis of time series involving coloured noise. For this, we introduce an algorithmic scheme for computing K_2 directly from a time series data. It is based on the GP algorithm and is an extention of the scheme we proposed earlier [14] for nonsubjective computation of D_2. The scheme is tested with a large number of standard chaotic systems and is found to give reliable results with respect to white as well as coloured noise contamination. Moreover, it can be applied to any arbitrary time series and provides an error estimate on the value of K_2^{sat} obtained. As examples from the real world, analysis of four astrophysical data sets have also been carried out.

Both D_2 and K_2 are very important measures to test nonlinearity in a time series data. But D_2 is more often chosen as the test statistic for hypothesis testing. This is because, computation of D_2 is much easier compared to that of K_2 and more importantly requires only less number of data points than for K_2. But as shown by

Radaelli *et al* [4], K_2 could be a more decisive measure for some data sets, especially those involving coloured noise. This is because, while coloured noise gives a well-saturated value for D_2 as a function of M [17] just like a chaotic system, the value of $K_2 \to 0$ as $M \to \infty$. But in order to use K_2 as a test statistic, it is important to have a nonsubjective approach for its computation so that, the same conditions are maintained in the algorithm for both the data and the surrogates, as is done in our scheme.

Another advantage of our computational scheme is that it can compute both D_2 and K_2 efficiently from a time series, thus enabling a more complete analysis of the data using the two complementary measures of low-dimensional chaos. Moreover, surrogate analysis can be performed with both D_2 and K_2 as discriminating statistic, which is essential to confirm the presence of low-dimensional chaos. The scheme presented here could be useful in this regard.

Acknowledgements

KPH and GA acknowledge the computing facilities in IUCAA, Pune.

References

[1] E Ott, *Chaos in dynamical systems* (Cambridge University Press, New York, 1993)
[2] S Neil Rasband, *Chaotic dynamics of nonlinear systems* (Wiley, New York, 1997)
[3] P Szepfalusy and G Gyorgyi, *Phys. Rev.* **A33**, 2852 (1996)
[4] S Radaelli, D Plewczynski and W M Macek, *Phys. Rev.* **E66**, 035202R (2002)
[5] K Urbanowicz and J A Holyst, *Phys. Rev.* **E67**, 046218 (2003)
[6] P Grassberger and I Proccacia, *Physica* **D9**, 189 (1983); *Phys. Rev. Lett.* **50**, 346 (1983)
[7] P Grassberger and I Proccacia, *Phys. Rev.* **A28**, 2591 (1983)
[8] K Judd, *Physica* **D56**, 216 (1992)
[9] J C Sprott and G Rowlands, *Int. J. Bifurcat. Chaos* **11**, 1865 (2001)
[10] Y Termonia, *Phys. Rev.* **A29**, 1612 (1984)
[11] J Theiler, *Phys. Rev.* **A36**, 4456 (1987)
[12] D Yu, M Small, R G Harrison and C Diks, *Phys. Rev.* **E61**, 3750 (2000)
[13] G Nolte, A Ziehe and K R Muller, *Phys. Rev.* **E64**, 016112 (2001)
[14] K P Harikrishnan, R Misra, G Ambika and A K Kembhavi, *Physica* **D215**, 137 (2006)
[15] J Theiler, *Phys. Lett.* **A155**, 480 (1991)
[16] R Misra, K P Harikrishnan, B Mukhopadhyay, G Ambika and A K Kembhavi, *Astrophys. J.* **609**, 313 (2004)
[17] A R Osborne and A Provenzale, *Physica* **D35**, 357 (1989)