Journal of Statistical Planning and Inference 193 (2018) 55–69

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Statistical moments of Gaussian kernel correlation sum and weighted least square estimator of correlation dimension and noise level

Zouhaier Dhifaoui

Department of family and community medicine, faculty of medicine of Sousse, Mohamed Karoui street, 4002 Sousse, Tunisia

A R T I C L E   I N F O

Article history:
Received 7 September 2016
Received in revised form 25 July 2017
Accepted 1 August 2017
Available online 12 August 2017

Keywords:
Gaussian kernel correlation integral
Gaussian kernel correlation sum
Non-central chi squared random variable
Polynomials chaos decomposition
Correlation dimension
Noise level

A B S T R A C T

In this paper, using non-central chi-squared distribution and polynomial chaos decomposition of a log-normal random variable, we derive the exact expressions for the covariance and variance of the Gaussian kernel correlation sum (Gkcs). The obtained results are combined with U-statistics theory and non-linear models theory to construct the exact confidence intervals for the correlation dimension and for the noise level in the case where deterministic time series are corrupted by additive Gaussian noise. The theoretical results are tested on two continuous chaotic dynamics corrupted by Gaussian noise for different values of signal to noise ratio (SNR). An application to a real data time series has been also conducted.

1. Introduction

In time series analysis the discrimination between deterministic and stochastic behaviors is an important goal, this discrimination allows an efficient prediction of time series which consist a crucial goal in statistical field, to reach this goal many methods have been developed and applied in different research fields (see for example Omidvarnia and Nasrabadi, 2008; Golestani et al., 2009; Gautama et al., 2004). The most known method relies on the use of the correlation dimension (see Small and Chi, 2003; Shirer et al., 1997; Tung et al., 1992 and some references therein) which is an efficient method for classification of data based on its behavior, using that this dimension characterizes the geometric properties of reconstructed attractor from time series for some embedding dimension, and then helps to identify the type of analyzed time series, for example a fractional value of the correlation dimension confirming the chaotic nature of the studied time series, also a divergence of correlation dimension when the embedding dimension increases indicates a stochastic behavior of time series, using that a stochastic process has an infinite degree of freedom and therefore it should show no tendency to unfold at any specific dimension, and in the deterministic case characterized by a finite degree of freedom this quantity saturates around a finite value when the embedding dimension increases. On the other hand, statistically, the correlation dimension is the most interesting using that it can be considered as a lower bound for the number of variables needed for modeling the dynamical system (for some applications example the readers can see Shiraj et al., 2005; Jason et al., 2008). The widely used method to estimate the correlation dimension is the Grassberger–Procaccia algorithm (Grassberger and Procaccia, 1983), this algorithm is used for behavioral study in different fields, such as in economic (Mohammad Ali et al., 2006; Hans-Walter, 2013), finance (Anna and Piotr, 2005; Dengyue, 2010), medicine (Berik et al., 2015; Loskutov and Mironyuk, 2007) and some other fields, and is based on the computation of correlation integral, which is a measure a spatial correlation of points in the attractor, on the using of the Heaviside step function. The principal disadvantage of this method is their sensitivity to the noise which is a

E-mail address: zouhaierdhifaoui@ymail.com.

http://dx.doi.org/10.1016/j.jspi.2017.08.001
0378-3758/

Notations

$\sum_{i,j}$:	$\sum_{i=1}^{N_m}\sum_{j=i+1}^{N_m}$
$\|.\|$:	Euclidean norm.
$\sim$:	Approximate.
X_d^2:	Central chi squared random variable with d degree of freedom.
$X_d^{'2}(\beta)$:	Non-central chi squared random variable with d degree of freedom and non-centrality parameter β.

principal characteristic of observed time series, using that the presence of noise strongly affects the correlation integral and then tends to increase the correlation dimension given in this case by the slope of linear regression of logarithmic transform of the correlation integral on logarithmic transform of bandwidth (Ying-Cheng and David, 1998). To overcome this problem, a Gaussian kernel function is used by Diks (1996) for estimating the correlation integral (details of Gkcs are remembering in Section 3) and permits to estimate the correlation dimension and the noise level in analyzed time series from a new scaling behavior of the correlation integral.

This paper, which is written to show the determination of exact confidence interval for correlation dimension and for the noise level in the case where a deterministic signal is corrupted by an additive Gaussian noise based on the use of Gkcs, is organized as follows: the introduction in Section 1, a motivation of this paper is presented in Section 2, Section 3 is devoted to theoretical concepts of Gaussian kernel correlation integral (Gkci) and their estimator (Gkcs) and to review the non-central chi squared distribution. In Section 4 we present a theoretical result followed by discussions and a proposed estimation method of correlation dimension and the noise level with the expression of their confidence intervals. Numerical tests are presented in Section 5 and an application to two real time series with interpretations is presented in Section 6, we conclude by Section 7.

2. Motivation

Different methods to estimate the correlation dimension are proposed based on the use of Gkcs. Diks (1996) proposes a method to estimate the correlation and entropy dimension and the noise level based on the using of Gkcs for range values of bandwidths smaller than 0.25, in this case the authors demonstrate the efficiency of their method for deterministic time series with up to 20% of noise. In Dejin et al. (2000b) the authors propose another method to estimate the noise level and the correlation dimension using Gkcs based on the suggested value of the largest bandwidth $h_{max} \leq 3\sigma_z$ where σ_z is the standard deviation of the noise part, this method is tested for high values of noise level and gives satisfactory results. On the other hand Nolte et al. (2001) develops a corrected estimator for correlation and entropy dimension and the noise level based on the difference of correlation dimensions computed for adjacent embedding dimensions, their method is efficient for a noise level equal to 50% and also it is efficient for moderately colored noise and gives satisfactory results, also, for non-Gaussian and a dynamical noise. A robust estimator to noise and outliers observations of correlation dimension is developed in Dhifaoui (2016) based on similarity between the evolution of Gkcs and that of modified Boltzmann sigmoidal function after logarithmic transformation, the simulation study indicates the robustness of the proposed estimator to the presence of different types of noise for high noise level, moreover, this estimator is also robust to the presence of 60% of outliers observations. But all these methods are not giving the statistical inference for the proposed estimator of correlation dimension and noise level, despite the fact that this problem is evoked in Svetlana et al. (2011) where the authors give the bootstrap distribution and the confidence interval for the correlation dimension estimated for one realization of chaotic time series based on the moving block and parametric bootstrap procedure combined with U-Statistics theory. The principal goal of this paper is to determine the confidence interval of correlation dimension and noise level in the case where a deterministic signal is corrupted by an additive Gaussian noise, based on the use of statistical properties of Gkcs.

3. Theoretical background

The correlation dimension D_2[1] is defined, when the limit exist, by

$$D_2 = \lim_{h\to 0} \frac{\log(C_m(h))}{\log(h)},$$

where $C_m(h)$ is the correlation integral (Grassberger and Procaccia, 1983) given by

$$C_m(h) = \mathbb{P}(\|\mathbf{X} - \mathbf{Y}\| \leq h),$$

where $\mathbf{X}$ and $\mathbf{Y}$ are independent and identically distributed reconstructed vectors.

[1] The subscript "2" refers to the case of the generic formula for the correlation dimension estimator originally discussed by Grassberger and Procaccia.