Chaos, Solitons and Fractals 93 (2016) 169–174

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

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

Robust to noise and outliers estimator of correlation dimension

Zouhaier Dhifaoui

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

ARTICLE INFO

Article history:
Received 1 September 2016
Revised 27 October 2016
Accepted 28 October 2016
Available online 2 November 2016

Keywords:
Gaussian kernel correlation integral
Correlation dimension
Modified Boltzmann sigmoidal function
Moving bloc bootstrap

ABSTRACT

The estimation of correlation dimension of continuous and discreet deterministic chaotic processes corrupted by an additive noise and outliers observations is investigated. In this paper we propose a new estimator of correlation dimension based on similarity between the evolution of Gaussian kernel correlation sum (Gkcs) and that of modified Boltzmann sigmoidal function (mBsf), this estimator is given by the maximum value of the first derivative of logarithmic transform of Gkcs against logarithmic transform of bandwidth, so the proposed estimator is independent of the choice of regression region like other regression estimators of correlation dimension. Simulation study indicates the robustness of proposed estimator to the presence of different types of noise such us independent Gaussian noise, non independent Gaussian noise and uniform noise for high noise level, moreover, this estimator is also robust to presence of 60% of outliers observations. Application of this new estimator with determination of their confidence interval using the moving block bootstrap method to adjusted closed price of S&P500 index daily time series revels the stochastic behavior of such financial time series.

1. Introduction

Financial fluctuations time series often tends to display chaos, statistical analysis of these property has formed a major area of research (the readers can see, for example, [1–4]). A prior knowledge of the processes governing the dynamics of these fluctuations facilitates the correct identification of their nature. Recently, the chaotic deterministic process has been proposed as an alternative for stochastic process for studying the behavior of some financial time series, using that such processes are characterized by some important properties such that they evolve according to specific rules, and don't explode because they are enclosed in a small space called attractor, and for modeling, they are efficient because they don't contain in their structure a non observation noise like a stochastic processes. Therefore, the detection of a deterministic behavior would mean an opportunity for hedgers, speculators as well as arbitrageurs to play the markets better. The important method to detect the determinism behavior in time series is based on the computation of correlation dimension measure developed by Grassberger and Procaccia [5] which is, in the first hand, the most interesting statistically, it is computed from the real data and serves to measure the dimension of the reconstructed attractor, also, it measures the complexity within observations and quantifies the spatial correlation between values which compose it, in the other hand the importance of the correlation dimension arises from the fact that the minimum number of variables to model a chaotic attractor is the smallest integer greater than it.

The estimation of correlation dimension necessitates the computation of correlation integral [5] defined by $C_m(h) = P(\|\mathbf{X} - \mathbf{Y}\| \leq h)$ where $\mathbf{X}$ and $\mathbf{Y}$ are independent and identically distributed reconstructed vectors from observed time series and m is the embedding dimension, in [5] the authors used the Heaviside step function H given by $H(x) = 1$ if $x \geq 0$ and 0 otherwise to estimate the correlation integral, and then estimate the correlation dimension D by the slope of linear part, for some small values of h, of $\log(\hat{C}_m(h))$ versus $\log(h)$ where $\hat{C}_m(h)$ is an estimator of correlation integral (see [5] for more details), Diks [6] uses the Gaussian kernel function given by $K(\mathbf{X},\mathbf{Y}) = \exp(-\|\mathbf{X} - \mathbf{Y}\|^2/4h^2)$ so as contributions from pairs with $\|\mathbf{X} - \mathbf{Y}\| > h$ do not vanish like Heaviside step function but are exponentially suppressed, using this kernel function the author demonstrates, based on the work of Ghez and Vaienti [7], that the correlation integral in the free noise case is approximated when $m \to \infty$ and $h \to 0$ by $\exp(-mK\delta_t)h^D$ (we review it in details in Section 2) and to estimate D he uses a nonlinear regression method on some range of h values.

The principal problem on the estimation of D is the choice of the region when we can do the regression, i.e., the region of values of bandwidth h where the logarithmic transform of $\hat{C}_m(h)$ is linear and parallel for consecutive values of m, this problem is evoked by researches (for surveys see [8,9]) especially when the deterministic time series is corrupted by noise which masks the scaling region at the small scale especially for high noise level, Diks in [6] pro-

E-mail address: zouhaierdhifaoui@ymail.com

http://dx.doi.org/10.1016/j.chaos.2016.10.017

pose a method to estimate the noise level, the correlation dimension and the entropy dimension based on the Gkcs using a rang of values of $h \leq 0.25$ and the authors demonstrates the efficiency of their method for deterministic time series with up to 20% of noise, in [10] the authors propose another method to estimate the noise level and the correlation dimension based on the suggested value of the largest bandwidth h_c ($h_c \leq 3\sigma$) where σ is the standard deviation of the noise part, this method is tested for a high values of noise level and give satisfactory results.

In this paper, we propose an estimator for correlation dimension based on similarity between the curve of logarithmic transform of Gkcs and that of mBsf, this estimator is independent on the choice of bandwidth h. This article is organized as follows: the introduction in Section 1, the Section 2 is devoted to review a theoretical concepts of Gaussian kernel correlation integral (Gkci) and their estimator Gkcs, also the simplified method to compute it. The proposed estimator is detailed in Section 3, the evaluation of such estimator by a simulation study is presented in Section 4 and their application to financial time series with determination of their non symmetric confidence interval using moving bloc bootstrap method is presented in Section 5, we conclude by Section 6.

2. Gaussian kernel correlation integral

2.1. Theoretical review

Suppose that we have a scalar time series $\{x_i\}, i = 1, \ldots, N_x$ sampled at equally spaced times $t_i = i\Delta_t$ where Δ_t is the sampling time interval and N_x denotes the length of the time series. The underlying attractor can be reconstructed using delay co-ordinates method [11], this reconstruction method consists in embedding the measured time series in an m-dimensional Euclidian space to create $N_m = N_x - (m-1)\tau$ delay state vectors $\{\mathbf{X}_i^m\}, i = 1, \ldots, N_m$ in terms of:

$$\mathbf{X}_i^m = [x_i, x_{i+\tau}, x_{i+2\tau}, \ldots, x_{i+(m-1)\tau}]^t \tag{1}$$

where τ is an integer referred to as a time lag (the delay time is given by $\delta_t = \tau\Delta_t$) and m is usually referred to as the embedding dimension.

The Gkci is used for estimating the correlation dimension, the entropy and the noise level for a contaminated time series, and can be summarized in two cases: the noise-free case and the case when the time series is corrupted by a Gaussian noise as follows (the reader can see [12] for more details):

1. In the case of noise-free scales, the Gkci $C_m(h)$ is defined as:

$$C_m(h) = \int\int p_m(\mathbf{X})p_m(\mathbf{Y})\exp\left(-\frac{\|\mathbf{X}-\mathbf{Y}\|^2}{4h^2}\right)d\mathbf{X}d\mathbf{Y}$$
$$\sim \exp\left(-mK\delta_t\right)\left(\frac{h}{\sqrt{m}}\right)^D, \text{ where } m \to +\infty \text{ and } h \to 0, \tag{2}$$

where D is the correlation dimension which characterizes the geometry of attractor in terms of its fractional dimension, K is the correlation entropy characterizes their time complexity and it quantifies the rate at which the distance between two initially nearby states increases under the dynamics, h is referred to as the bandwidth and p_m is a probability distribution function. The scaling behavior in Eq. (2) was first justified by Ghez et al. (the reader can see [13] Eq. (8) and Theorem 1, and also [7] Eq. (4.6)) and latter by Diks (see [6] Eq. (7)) with inclusion of the factor in which the m dependence was originally introduced by Frank et al. [14] to improve the convergence of correlation entropy K.

2. In the presence of Gaussian noise the distribution function becomes p_m^* which can be expressed in terms of a convolution between the underlying noise-free distribution function p_m and a normalized Gaussian distribution function p_m^g with standard deviation equal to σ, i.e.,

$$p_m^*(\mathbf{Y}) = \int p_m(\mathbf{X})p_m^g(\|\mathbf{Y}-\mathbf{X}\|)d\mathbf{X}$$
$$= \frac{1}{(\sigma\sqrt{2\pi})^m}\int p_m(\mathbf{X})\exp\left(\frac{-\|\mathbf{Y}-\mathbf{X}\|^2}{2\sigma^2}\right)d\mathbf{X} \tag{3}$$

accounts for noise effects in m-dimensional space [6,15], then the Gkci $C_m^*(h)$ in this case and using Eqs. (2) and (3) have the corresponding scaling law (when $\sqrt{h^2+\sigma^2} \to 0$ and $m \to +\infty$):

$$C_m^*(h) = \int\int p_m^*(\mathbf{X})p_m^*(\mathbf{Y})\exp\left(-\frac{\|\mathbf{Y}-\mathbf{X}\|^2}{4h^2}\right)d\mathbf{X}d\mathbf{Y} \tag{4}$$

$$= \left(\frac{h^2}{h^2+\sigma^2}\right)^{\frac{m}{2}}\int\int p_m(\mathbf{X})p_m(\mathbf{Y})\exp\left(-\frac{\|\mathbf{X}-\mathbf{Y}\|^2}{4(h^2+\sigma^2)}\right)d\mathbf{X}d\mathbf{Y}$$
$$\simeq \alpha\left(\frac{h^2}{h^2+\sigma^2}\right)^{\frac{m}{2}}\exp(-mK\delta_t)\left(\frac{h^2+\sigma^2}{m}\right)^{\frac{D}{2}}, \tag{5}$$

where α is a normalized constant.
The parameter σ is referred to as the noise level defined by $\sigma = \sigma_n/\sigma_x = \sigma_n/\sqrt{\sigma_d^2+\sigma_n^2}$ where σ_x, σ_d and σ_n are the standard deviation of the input noisy signal $\{x_i\}$, underlying clean component $\{d_i\}$ and the Gaussian noise part $\{n_i\}$, where in total signal $x_i = d_i + n_i$, d_i and n_i are assumed to be statistically independent.

2.2. Estimator and simplified algorithm

In the case of discrete sampling and assuming that the vector points on the attractor are dynamically independently distributed according to p_m^* and using an average over delay vectors to replace the integral over the vector distributions in Eq. (4), consequently, $C_m^*(h)$ can be computed as:

$$\hat{C}_m^{\mathbf{g}}(h) = \frac{1}{N_m(N_m-1)}\sum_{i=1}^{N_m}\sum_{j\neq i}^{N_m}\exp\left(-\frac{\|\mathbf{Y}_i^m-\mathbf{Y}_j^m\|^2}{4h^2}\right). \tag{6}$$

For estimating D, K and σ, $\hat{C}_m^{\mathbf{g}}(h)$ is computed for a series of discrete bandwidth values $\{h_k\}, k = 0, 1, 2, \ldots, N_b$ and fitting to the scaling relation given by Eq. (5) using nonlinear least squares method (see [16] for theoretical review of this method and associated computation algorithms).

The implementation of Gaussian kernel algorithm (GKA) [6] developed to compute $\hat{C}_m^{\mathbf{g}}(h)$ needs the transformation of the input signal $\{x_i\}$ on the new signal ν_i according to the formula:

$$\nu_i = \frac{x_i - \bar{x}}{\sigma_x}, \ i = 1, \ldots, N_x \tag{7}$$

where $\bar{x}$ and σ_x denotes the mean and the standard deviation of the input signal $\{x_i\}$, under this transformation the noise effect is described by the distribution function p_m^g and the standard deviation of the noise part is σ, accordingly, the delay state vectors are reconstructed by replacing $\{x_i\}$ with $\{\nu_i\}$.

A direct computation of $\hat{C}_m^{\mathbf{g}}(h)$ is characterized by a computational complexity of order $\mathcal{O}(N_m^2 \times N_b)$, to eliminate this highly time complexity Yu et al. [10] has developed an efficient algorithm based on a simplification of expression given by Eq. (6) by showing