Draft: Symbolic Correlation Integral. Getting Rid of the Proximity Parameter

M Vitoria Caballero Mariano Matilla-García Manuel Ruiz Marín.

Abstract

In this paper we introduce the symbolic correlation integral $SC(m)$, which avoids the noisy parameter ε of the classical correlation integral defined by Grassberger-Procaccia. Moreover we provide the asymptotic distribution of $SC(m)$ under the null of i.i.d.. With a MonteCarlo simulation we show the size and the power performance of the new test under linear and nonlinear processes.

1 Introduction

Correlation integral has been widely used in finance and macroeconomics basically for testing serial independence. The most relevant contribution is the well-known non-parametric BDS test [5], which is precisely based on the fact that correlation integral factorizes when the elements of a time series are i.i.d. (independent and identically distributed). The BDS test of Brock, Dechert, Sheinkeman and LeBaron, is now an important part of most standard econometric data analysis software packages. Interestingly, correlation integral has also been used [9] to detect potential nonlinear causal relationships between time series. Correlation integral, introduced by [8], and initially designed for measuring the fractal dimension of chaotic data, currently constitutes an important tool also in physics and natural sciences for analyzing some time series' properties (see for example [4], [17] [7]; and [14], for the latest contributions).

[12] have shown that the BDS test is highly suitable for economic time series analysis. Indeed, the BDS test has been extensively used to analyze nonlinear structures in many economic time series: Financial market data like stock market returns, exchange rate returns and daily oil production were studied, among others, by [13], [10] and [16], respectively. The correlation integral based statistic has also been used to test for nonlinearity in many other economic domains such as those related with GDP, unemployment, agricultural prices, and so on.

By construction, the BDS test depends on the proximity parameter, which is a central element in the definition of correlation integrals, and therefore one can easily obtained two opposite statistical conclusions depending on the initial parameter choice. This potential outcome is as serious problem for practitioners of correlation based tests. In addition, there have been several critical comments to the BDS test ([18], [6], [3], among others) regarding a larger over-estimation of the right level of the test, hence tending to report a high false alarm rate. This can be explained, in part, because of the non trivial impact of several ex ante decisions about the proximity parameter that are required to be taken in order to use the test. Unfortunately there is very limited statistical guidance (most of them based on massive Monte Carlo experiments) about its selection, being likely to choose inappropriate values. being a notable exception is the empirical study conducted by [11].

In this paper we explicitly propose a new statistic based on correlation integral, but that does not depend, by definition, on the proximity parameter. Instead, we rely on symbolic dynamics that has been proved to be an interesting approach to test for independence and causality, among others.

2 Definitions and notation

Let $\{x_t\}_{t\in I}$ be a real-valued time series from a strictly stationary stochastic process of real random variables, where I is a set of time indexes of length T. For a positive integer $m \geq 2$ we denote by S_m the symmetric group of order $m!$, that is the group formed by all the permutations of length m. Let $\pi_i = (i_1, i_2, \ldots, i_m) \in S_m$. We will call an element π_i in the symmetric group S_m a symbol. The positive integer m is usually known as *embedding dimension*.

Now we define an ordinal pattern for a symbol $\pi_i = (i_1, i_2, \ldots, i_m) \in S_m$ at a given time $t \in I$. To this end we consider that the time series is embedded in an m-dimensional space as follows:

$$\overline{x}_t = (x_t, x_{t+1}, ..., x_{t+(m-1)}).$$

Then we say that $\overline{x}_t$ is of π_i-type if and only if $\pi_i = (i_1, i_2, \ldots, i_m)$ is the unique symbol in the group S_m satisfying the two following conditions:

$$(a) \quad x_{t+i_1} \leq x_{t+i_2} \leq \cdots \leq x_{t+i_m}, \text{ and}$$

$$(b) \quad i_{s-1} < i_s \text{ if } x_{t+i_{s-1}} = x_{t+i_s}$$

Condition (b) guaranties uniqueness of the symbol π_i. This is justified if the values of x_t have a

continuous distribution so that equal values are very uncommon, with a theoretical probability of occurrence of 0.

Notice that $\{\overline{x}_t\}_{t\in\widehat{I}}$ is a vectorial time series where the set $\widehat{I}$ has length $n = T - m + 1$. Each $\overline{x}_t$ is called $m-$history.

Notice that even in the case in which the time series $\{x_t\}_{t\in I}$ is i.i.d., the vectorial time series $\{\overline{x}_t\}_{t\in\widehat{I}}$ is $m-$dependent due to the overlapping that each $m-$history has with the next $m-1$ $m-$histories as seen below:

$$\begin{array}{lllllll}
\overline{x}_t = (x_t & x_{t+1} & x_{t+2} & \dots & x_{t+m-1}) & & \\
\overline{x}_{t+1} = (x_{t+1} & x_{t+2} & \dots & x_{t+m-1} & x_{t+m}) & & \\
\overline{x}_{t+2} = (x_{t+2} & \dots & x_{t+m-1} & x_{t+m} & x_{t+m+1}) & & \\
\vdots & & & & & & \\
\overline{x}_{t+m-1} = (x_{t+m-1} & x_{t+m} & x_{t+m+1} & \dots & x_{t+2(m-1)}) & &
\end{array}$$

We will say that $\overline{x}_i$ and $\overline{x}_j$ overlap when $|j-i| < m$. If this is the case it will be denoted by $j \in N(i)$.

Next we define the symbolization map

$$s : \{\overline{x}_t\}_{t\in\widehat{I}} \longrightarrow S_m$$

defined by $s(\overline{x}_t) = \pi$ if and only if $\overline{x}_t$ is of $\pi-$type. Notice that the symbolization map s transforms the vectorial time series of m-histories in a sequence of symbols.

Next we introduce all probabilities in the symbols space that will be needed to estimate the variance of the new statistic.

Given two m-histories $\overline{x}_i$ and $\overline{x}_j$ we will denote by $p_{ij}^{\pi\pi}$ the probability of $\overline{x}_i$ of being of the same symbol π-type than $\overline{x}_j$. Similarly we will denote by $p_{ij}^{\pi\delta}$ the probability of $\overline{x}_i$ of being of π-type and $\overline{x}_j$ of δ-type.

Notice that in the case in which the time series $\{x_t\}_{t\in I}$ is i.i.d. and the m-histories $\overline{x}_i$ and $\overline{x}_j$ do not overlap (i.e. $|j-i| > m$ or equivalently $j \notin N(i)$) it follows that $p_{ij}^{\pi\pi} = p_{ij}^{\pi\delta} = (1/m!)^2$. This does not occur when $|j-i| \leq m$ and therefore the $m! \times m!$ probability matrix $PM(|j-i|) = (p_{ij}^{\pi\delta})$ have to be computed for all $\pi, \delta \in S_m$.

For example, when $m = 3$ we have computed the different probability matrixes when $|j-i| = 1$, $|j-i| = 2$ and $|j-i| \geq 3$:

$$PM(1) = 1/3! \begin{pmatrix} 0.25 & 0.25 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0.25 & 0.25 & 0 & 0.5 \\ 0.25 & 0.5 & 0 & 0 & 0 & 0.25 \\ 0.5 & 0.25 & 0 & 0 & 0 & 0.25 \\ 0 & 0 & 0.25 & 0.5 & 0 & 0.25 \\ 0 & 0 & 0.5 & 0.25 & 0 & 0.25 \end{pmatrix}$$

$$PM(2) = 1/3! \begin{pmatrix} 0.05 & 0.05 & 0.15 & 0.3 & 0.15 & 0.3 \\ 0.15 & 0.15 & 0.2 & 0.15 & 0.2 & 0.15 \\ 0.05 & 0.05 & 0.15 & 0.3 & 0.15 & 0.3 \\ 0.15 & 0.15 & 0.2 & 0.15 & 0.2 & 0.15 \\ 0.3 & 0.3 & 0.15 & 0.05 & 0.15 & 0.05 \\ 0.3 & 0.3 & 0.15 & 0.05 & 0.15 & 0.05 \end{pmatrix}$$

Also if $k = |j - i| \geq 3$ then the m-histories do not overlap and therefore the probability matrix remains as

$$PM(k) = 1/3! \begin{pmatrix} 1/3! & 1/3! & 1/3! & 1/3! & 1/3! & 1/3! \\ 1/3! & 1/3! & 1/3! & 1/3! & 1/3! & 1/3! \\ 1/3! & 1/3! & 1/3! & 1/3! & 1/3! & 1/3! \\ 1/3! & 1/3! & 1/3! & 1/3! & 1/3! & 1/3! \\ 1/3! & 1/3! & 1/3! & 1/3! & 1/3! & 1/3! \\ 1/3! & 1/3! & 1/3! & 1/3! & 1/3! & 1/3! \end{pmatrix}$$

Now, we define the indicator variable

$$I(s(\overline{x}_t), s(\overline{x}_s)) = \begin{cases} 1 & \text{if } s(\overline{x}_t) = s(\overline{x}_s) \\ \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This indicator function, $I(s(\overline{x}_i), s(\overline{x}_j))$ is a bernoulli variable with probability of success

$$p_{ij} = \sum_{\pi \in S_m} p_{ij}^{\pi\pi}.$$

For four time indexes i, j, t, s we define the following probabilities

$$p(i, j, t, s) = p(I(s(\overline{x}_i), s(\overline{x}_j)) = 1, I(s(\overline{x}_t), s(\overline{x}_s)) = 1).$$

To calculate all of these probabilities under certain scenarios of overlapping will be essential in order to compute the variance of our statistic under the null of i.i.d.. When the time series $\{x_t\}_{t\in I}$ is i.i.d., the probabilities $p(i, j, s, t)$ can be computationally[1] calculated for small values of m, $m \leq 5$. For greater values of m there is need of a super-computer to obtain these probabilities. In this paper we have computed $p(i, j, t, s)$ for $m \leq 5$. To this end, notice that if the time series $\{x_t\}_{t\in I}$ is i.i.d., without loss of generality we may assume that $i = 1$ and $s > t$. More concretely, next we show all possible combinations of time indexes $1, j, t, s$ in order to compute the probabilities $p(1, j, t, s)$ that will be of interest. This is the case for $j > 1$ and $1 \leq t \leq m$. In order for the reader to better understand this point, we will show the value of the probabilities for all of these possible scenarios when $m = 3$ for the four m-histories under consideration. Notice that we will only show the first four decimal places.

1. $j, s \notin N(1)$ and $s \in N(j)$. Then

$$p(1, j, 1, s) = tracePM(s - j) \cdot \frac{1}{m!}$$

For $m = 3$ we obtain the following probabilities:

$$p(1, j, 1, j + 1) = 5/360 \quad p(1, j, 1, j + 2) = 7/360$$

for $j \geq 4$.

2. $j \in N(1)$ and $s \in N(j)$ but $s \notin N(1)$. Then

$$p(1, j, 1, s) = \sum_{\pi\in S_m} p_{1j}^{\pi\pi} \cdot p_{js}^{\pi\pi}$$

For $m = 3$ we obtain the following probabilities:

$$p(1, 2, 1, 4) = \textstyle\sum_{\pi\in S_m} p_{12}^{\pi\pi} \cdot p_{24}^{\pi\pi} = 0.0250$$
$$p(1, 3, 1, 4) = \textstyle\sum_{\pi\in S_m} p_{13}^{\pi\pi} \cdot p_{34}^{\pi\pi} = 0.0250$$
$$p(1, 3, 1, 5) = \textstyle\sum_{\pi\in S_m} p_{13}^{\pi\pi} \cdot p_{35}^{\pi\pi} = 0.0950$$

[1]Codes in matlab are available upon request

3. $j, s \in N(1)$ with $j \neq s$. It is not possible to compute $p(1, j, 1, s)$ with the probability matrices. Therefore we will make use of a matlab code to compute them. When $m = 3$, there is only one possibility

$$p(1, 2, 1, 3) = 1/60$$

4. If the m-histories at $1, j, s$ do not overlap then

$$p(1, j, 1, s) = \frac{1}{m!^2}.$$

5. $j \notin N(s)$. In this case we have four possibilities

   (a) If $j \notin N(1)$ and $s \in N(t)$ then

$$p(1, j, t, s) = \frac{1}{m!} tracePM(s - t)$$

   For m=3 and every $j > 4$ we have

$$p(1, j, 2, 3) = \sum_{\delta \in S_m} p_{23}^{\delta\delta} = \frac{5}{360}$$

$$p(1, j, 2, 4) = \sum_{\delta \in S_m} p_{24}^{\delta\delta} = \frac{7}{360}$$

$$p(1, j, 3, 4) = \sum_{\delta \in S_m} p_{34}^{\delta\delta} = \frac{5}{369}$$

   (b) If $j \in N(1)$ and $s \notin N(t)$ then

$$p(1, j, t, s) = \frac{1}{m!} tracePM(j - 1).$$

   If m=3 and every $s > 4$ we have

$$p(1, 2, 2, s) = \sum_{\pi \in S_m} p_{12}^{\pi\pi} = \frac{5}{360}$$

$$p(1, 3, 2, s) = \sum_{\pi \in S_m} p_{13}^{\pi\pi} = \frac{7}{360}$$

$$p(1, 2, 3, s) = \sum_{\pi \in S_m} p_{12}^{\pi\pi} = \frac{5}{360}$$

$$p(1, 3, 3, s) = \sum_{\pi \in S_m} p_{13}^{\pi\pi} = \frac{7}{360}$$

(c) If $j \notin N(1)$ and $s \notin N(t)$ it follows that

$$p(1,j,t,s) = \frac{1}{m!}$$

(d) If $j \in N(1)$ and $s \in N(t)$, but $s \notin N(1)$ then

$$p(1,j,t,s) = tracePM(j-1) \cdot tracePM(s-t)$$

.

When $m = 3$ we obtain

$$p(1,2,3,5) = \sum_{\pi \in S_m} p_{12}^{\pi\pi} \cdot \sum_{\delta \in S_m} p_{35}^{\delta\delta} = \frac{5}{360}$$

6. When $s \in N(j)$ the probabilities $p(1,j,t,s)$ must be computed. We distinguish several cases:

(a) $j, s \notin N(1) \cup N(t)$. In this case the probability $p(1,j,t,s)$ can be computed with the probability matrices PM

$$p(1,j,t,s) = \sum_{\pi \in S_m} \sum_{\delta \in S_m} p_{1t}^{\pi\delta} p_{js}^{\pi\delta}$$

When $m = 3$ we have that:

(b) $s \in N(t)$, but $s \notin N(1)$

When $m = 3$ we obtain

$$\begin{aligned}
t = 2, s = 4, j = 5 \quad & p(1,5,2,4) = 0.0194 \\
t = 2, s = 4, j = 6 \quad & p(1,6,2,4) = 0.0237 \\
t = 3, s = 4, j = 6 \quad & p(1,6,3,4) = 0.0174 \\
t = 3, s = 5, j = 6 \quad & p(1,6,3,5) = 0.0237 \\
t = 3, s = 5, j = 7 \quad & p(1,7,3,5) = 0.0210
\end{aligned}$$

(c) $s \in N(1)$, but $j \notin N(t)$

If $m = 3$, $t = 2, s = 3, j = 5$,

$$p(1,5,2,3) = 0.0183.$$

(d) $j \in N(t)$, but $\jmath \notin N(1)$ When $m = 3$ we have that

$$\begin{aligned}
t = 2, s = 5, j = 4 \quad & p(1,4,2,5) = 0.0651 \\
t = 2, s = 6, j = 4 \quad & p(1,4,2,6) = 0.0191 \\
t = 3, s = 6, j = 4 \quad & p(1,4,3,6) = 0.0354 \\
t = 3, s = 6, j = 5 \quad & p(1,5,3,6) = 0.0191 \\
t = 3, s = 7, j = 5 \quad & p(1,5,3,7) = 0.0351
\end{aligned}$$

(e) $j \in N(1)$, but $s \notin N(t)$ For $m = 3$, $t = 2, s = 5, j = 3, p(1, 3, 2, 5) = 0.0123$

(f) $j \in N(t)$ but $j \notin N(1)$ and $s \in N(t)$ but $s \notin N(1)$.

When $m = 3$ it follows that

$$\begin{aligned} t = 2, s = 4, j = 4 \quad & p(1, 4, 2, 4) = 0.0028 \\ t = 3, s = 4, j = 4 \quad & p(1, 4, 3, 4) = 0.0028 \\ t = 3, s = 5, j = 5 \quad & p(1, 5, 3, 5) = 0.0123 \\ t = 3, s = 4, j = 5 \quad & p(1, 5, 3, 4) = 0.0183 \\ t = 3, s = 4, j = 4 \quad & p(1, 4, 3, 5) = 0.0123 \end{aligned}$$

(g) $j \in N(1)$ and $j \in N(t)$ but $s \notin N(1)$.

If $m = 3$ we obtain

$$\begin{aligned} t = 2, s = 4, j = 2 \quad & p(1, 2, 2, 4) = 0.0028 \\ t = 3, s = 4, j = 3 \quad & p(1, 3, 3, 4) = 0.0028 \\ t = 3, s = 5, j = 3 \quad & p(1, 3, 3, 5) = 0.0123 \\ t = 3, s = 4, j = 2 \quad & p(1, 2, 3, 4) = 0.0028 \\ t = 2, s = 4, j = 3 \quad & p(1, 3, 2, 4) = 0.0183 \end{aligned}$$

(h) $j \in N(1)$ and $s \in N(t)$.

For $m = 3$

$$\begin{aligned} t = 2, s = 3, j = 4 \quad & p(1, 4, 2, 3) = 0.0028 \\ t = 2, s = 3, j = 2 \quad & p(1, 2, 2, 3) = 0.0167 \\ t = 2, s = 3, j = 3 \quad & p(1, 3, 2, 3) = 0.0167 \end{aligned}$$

3 The Symbolic Correlation Integral

In this section we will define the concept of symbolic correlation integral and we will derive its asymptotic distribution under the null of i.i.d for the time series $\{x_t\}_{t \in I}$.

Definition 1. *The symbolic correlation integral of a set $X \subset \mathbb{R}^m$ distributed according to a measure μ is defined as*

$$SC(X) = \int \int I(s(\overline{x}), s(\overline{y})) d\mu(\overline{x}) d\mu(\overline{y}) = E(I \circ (s \times s)(X)).$$

That is, symbolic correlation integral measures the probability that two elements in X have the same ordinal pattern.

We are interested in symbolic correlation integral of an embedded vectorial time series. In this setting, for $m \geq 2$ and the finite m-embedded vectorial time series $\{\overline{x}_t\}_{t=1}^n$, an estimation of the symbolic correlation integral of the set $\{\overline{x}_t\}_t$, is given by

$$SC_n = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(s(\overline{x}_i), s(\overline{x}_j))$$

Notice that the statistic SC_n is an $U-$statistic where $I \circ (s \times s)$ is its symmetric kernel. It is straightforward to check that SC_n satisfies the conditions of the theorem of Aaronson et al. (1996) and thus it satisfy the U-statistic ergodic theorem. Therefore we have that

$$\lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(s(\overline{x}_i), s(\overline{x}_j)) = \int \int I(s(\overline{x}), s(\overline{y}))d\mu(\overline{x})d\mu(\overline{y}). \tag{2}$$

3.1 Asymptotic Distribution of the Symbolic Correlation Integral

In this section we will compute the asymptotic distribution of the Symbolic Correlation Integral under the null of i.i.d for a time series. The proof of the main theorem of this section strongly relies on the following result in Denker and Keller (1983).

Theorem 3.1 (Denker and Keller (1983))**.** *Let $\{X_t\}_{t\geq 1}$ be absolutely regular with mixing coefficients satisfying $\sum_{k=1}^{\infty} \beta_k; \delta/(2+\delta) < \infty$ and let $h : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric kernel satisfying $\sup_{i<j} E(|h(X_i, X_j)|^{2+\delta}) < \infty$. Let*

$$U_n(h) = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} h(X_i, X_j)$$

be a U-statistic with expected value θ and define $h_1(x) = E(h(x, Y)) - \theta$. Then

$$\sqrt{n}(U_n(h) - \theta)$$

is asymptotically distributed as $N(0, 4\sigma^2)$ where $\sigma^2 = \lim_{n\to\infty} \frac{1}{n} Var(\sum_{j=1}^{n} h_1(X_j))$, always $\sigma^2 \neq 0$.

In order to state our main result we define

$$I_1(s(\overline{x}_1)) = E\left[I(s(\overline{x}_1), s(\overline{x}))\right] - \frac{1}{m!}.$$

Theorem 3.2. *Let $\{x_t\}$ be an i.i.d strictly stationary real valued time series and let $\{\overline{x}_t\}$ be the embedded time series of $m-$histories. Then*

$$\sqrt{n}\left(SC_n - \frac{1}{m!}\right)$$

converges in distribution to $\mathbf{N}(0, 4\sigma^2)$ *where*

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} Var\left(\sum_{j=1}^{n} I_1(s(\overline{x}_j))\right),$$

always $\sigma \neq 0$.

Proof. Notice that if the time series is i.i.d. the vectorial time series $\{\overline{x}_t\}$ is absolutely regular with mixing coefficients satisfying that $\beta_s = 0$ for all $s > m$.

Next we want to compute the expected value of the statistic symbolic correlation integral, SC_n, for each n, when $\{x_t\}_{t\in I}$ is i.i.d.

$$\begin{aligned} E[SC_n] &= \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[I(s(\overline{x}_i), s(\overline{x}_j))] = \\ &= \frac{2}{n(n-1)} \left(\sum_{i=1}^{n-m+1} \sum_{j=i+1}^{i+m-1} E[I(s(\overline{x}_i), s(\overline{x}_j))] + \sum_{i=n-m+1}^{n-1} \sum_{j=i+1}^{n} E[I(s(\overline{x}_i), s(\overline{x}_j))] + \right. \\ &\quad \left. + \sum_{i=1}^{n-m} \sum_{j=i+m}^{n} E[I(s(\overline{x}_i), s(\overline{x}_i))] \right) \end{aligned} \tag{3}$$

After a few calculations we obtain

$$E[SC_n] = \frac{2}{n(n-1)} \left[\frac{1}{m!} \sum_{i=1}^{n-1} \sum_{j=i+1}^{i+m-1} trace(PM(j-i)) + \sum_{i=1}^{n-1} \sum_{j=i+m}^{n} \frac{1}{m!} \right] \tag{4}$$

Therefore

$$\lim_{n\to\infty} E[SC_n] = \frac{1}{m!}. \tag{5}$$

Therefore by Theorem3.1 it follows that

$$\sqrt{n}\left(SC_n - \frac{1}{m!}\right)$$

converges in distribution to $\mathbf{N}(0, 4\sigma^2)$ where

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} Var\left(\sum_{j=1}^{n} I_1(s(\overline{x}_j))\right),$$

always $\sigma \neq 0$, as desired. □

Nevertheless the estimation of the variance of SC is more complicated. To this end we relies on the theory of $U-$statistics and the Hoeffding decomposition (Hoeffding, 1948, 1961).

We want to estimate

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} var\left(\sum_{j=1}^{n} I_1(s(\overline{x_j}))\right) = E(I_1(s(\overline{x_1}))) + 2\sum_{t=2}^{m} E(I_1(s(\overline{x_1}))I_1(s(\overline{x_t}))) \tag{6}$$

provided the sum converges absolutely.

The first term of (6) is

$$E\left(I_1(s(\overline{x}_1))\right)^2 = var(I_1(s(\overline{x}_1)))$$

because $E[I_1(s(\overline{x}_1))] = 0$.

A estimation of $I_1(s(\overline{x}_1))$ is given by

$$\frac{1}{n-1}\sum_{i=2}^{n} I(s(\overline{x}_1), s(\overline{x}_i))$$

and therefore an estimation of $E\left(I_1(\overline{x}_1)\right)^2$ is given by

$$\frac{1}{(n-1)^2}\left[\sum_{i=2}^{n} var(I(s(\overline{x}_1), s(\overline{x}_i))) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} cov(I(s(\overline{x}_1), s(\overline{x}_i)), I(s(\overline{x}_1), s(\overline{x}_j)))\right]$$

where

$$var(I(s(\overline{x}_1), s(\overline{x}_i))) = \frac{1}{m!} trace(PM(i-1))\left(1 - \frac{1}{m!} trace(PM(i-1))\right)$$

when $i \in N(1)$.

In the case $i > m$,

$$var(I(s(\overline{x}_1), s(\overline{x}_i))) = \frac{1}{m!}\left(1 - \frac{1}{m!}\right)$$

On the other hand, following the notation in Section 2 it follows that

$$cov(I(s(\overline{x}_1), s(\overline{x}_i)), I(s(\overline{x}_1), s(\overline{x}_j))) = p(1, i, 1, j) - p_{1i}p_{1j}$$

is different from zero when $j \in N(i)$.

The second term of (6) is

$$2\sum_{t=2}^{m} cov(I_1(s(\overline{x}_1)), I_1(s(\overline{x}_t)))$$

The covariances $cov(I_1(s(\overline{x}_1)), I_1(s(\overline{x}_t)))$ can be estimated by

$$\frac{1}{(n-1)(n-t)}\sum_{i=2}^{n}\sum_{s=t+1}^{n} E\left(I(s(\overline{x}_1), s(\overline{x}_i)I(s(\overline{x}_t), s(\overline{x}_s))\right) - E\left(I(s(\overline{x}_1), s(\overline{x}_i))\right) E\left(I(s(\overline{x}_t), s(\overline{x}_s))\right) =$$
$$= \frac{1}{(n-1)(n-t)}\sum_{i=2}^{n}\sum_{s=t+1}^{n} p(1, i, t, s) - p_{1i}p_{ts}$$

Moreover all cases with interest are those described in Section 2, otherwise the covariances are zero.

3.2 The case of non-overlapping m-histories

If one consider the the non-overlapping m-histories:

$$\overline{x}_t = (x_{(t-1)m+1}, x_{(t-1)m+2}, ..., x_{(t-1)m+m}) \tag{7}$$

for $t = 1, 2, ..., [T/m]$ where $[\cdot]$ denotes the integer part of a real number, we can state the following theorem as a direct application of the Central Limit Theorem

Theorem 3.3. *Let $\{x_t\}_{t=1}^{T}$ be an i.i.d time series and consider the sequence of non-overlapping m- histories $\{\overline{x}_t\}_{t=1}^{k}$, where $k = [T/m]$ as given in (7). Then the statistic*

$$\frac{SC_k - \frac{1}{m!}}{\sqrt{\frac{2}{k(k-1)}\left(\frac{1}{m!}\left(1 - \frac{1}{m!}\right)\right)}}$$

asymptotically follows a $N(0,1)$ distribution.

Proof. In this case the statistic symbolic integral correlation can be written as

$$SC_k = \frac{2}{k(k-1)}\left[\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} I_{ij}\right]$$

where $k = [T/m]$ and I_{ij} are independent random variables bernoulli of parameter $\dfrac{1}{m!}$, then

$$\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} I_{ij} \sim \mathbf{B}\left(\frac{k(k-1)}{2}, \frac{1}{m!}\right)$$

is Binomial distributed and hence SC_k is distributed as

$$\frac{2}{k(k-1)}\mathbf{B}\left(\frac{k(k-1)}{2}, \frac{1}{m!}\right).$$

Then by the Normal approximation of the binomial distribution we obtain that

$$\frac{SC_k - \frac{1}{m!}}{\sqrt{\frac{2}{k(k-1)}\left(\frac{1}{m!}\left(1 - \frac{1}{m!}\right)\right)}}$$

asymptotically follows a $N(0,1)$ distribution, as desired. □

4 Finite-Sample behavior

To sshow the size behavior of the new test estatistic we have considered symetric and non-symetric distributions. More concretely:

1. Gaussian distribution with zero mean and unit variance, $\mathbf{N}(0,1)$.
2. Uniform distribution on the $(0,1)$ interval, $\mathbf{U}(0,1)$.
3. Chi-square distribution with 4 degree of freedom, χ^2_4.
4. Student's $t-$distribution with 4 degree of freedom, t_4.

Each process has been repeated 20000 times and the proportion of rejection of the i.i.d. null has been calculated using a nominal level of 5%. We have considered the values for m to be 3,4 and 5, and $T = 250, 500, 1000, 1500, 2000, 2500$ and 3000.

m=3	$T=250$	$T=500$	$T=1000$	$T=1500$	$T=2000$	$T=2500$	$T=3000$
$N(0,1)$	0.0455	$0,0474$	$0,0482$	$0,0501$	$0,0487$	$0,0511$	$0,0491$
χ^2_4	$0,0474$	$0,0488$	$0,0502$	$0,0484$	$0,0491$	$0,0509$	0.0486
$U(0,1)$	0.0495	$0,0511$	$0,0477$	$0,0501$	$0,0492$	$0,0514$	$0,0517$
t_4	0.0462	0.0472	0.0522	0.0526	0.0493	0.0515	0.0502

m=4	$T=250$	$T=500$	$T=1000$	$T=1500$	$T=2000$	$T=2500$	$T=3000$
$N(0,1)$	0.04205	$0,0413$	$0,04395$	$0,0416$	$0,0422$	$0,043$	$0,0444$
χ^2_4	$0,041$	$0,042$	$0,0425$	$0,0451$	$0,0465$	$0,045$	0.045
$U(0,1)$	0.042	$0,042$	$0,0425$	$0,042$	$0,043$	$0,043$	$0,0435$
t_4	0.040	0.0423	0.0427	0.0426	0.0429	0.0425	0.0435

m=5	$T=250$	$T=500$	$T=1000$	$T=1500$	$T=2000$	$T=2500$	$T=3000$
$N(0,1)$	0.039	0.0426	0.0454	0.0455	0.0467	0.0475	0.476
χ^2_4	0.040	0.0453	0.0454	0.045	0.0465	0.0472	0.0462
$U(0,1)$	0.038	0.0451	0.0451	0.0492	0.0473	0.0493	0.0471
t_4	0.0399	0.0454	0.0457	0.0464	0.0479	0.0476	0.0483

In order to show the power performance of SC, we have considered the following data generating processes (DGPs) because of its rich nonlinear variety. The models are the following:

DGP 1 $x_t = \epsilon_t + 0.8\epsilon_{t-1}^2$,

DGP 2 $x_t = \epsilon_t + 0.8\epsilon_{t-2}^2$,

DGP 3 $x_t = \epsilon_t + 0.8\epsilon_{t-3}^2$,

DGP 4 $x_t = \epsilon_t + 0.8\epsilon_{t-1}^2 + 0.8\epsilon_{t-2}^2 + 0.8\epsilon_{t-3}^2$,

DGP 5 $x_t = |x_{t-1}|^{0.8} + \epsilon_t$,

DGP 6 $x_t = sign(x_{t-1}) + \epsilon_t$,

DGP 7 $x_t = 0.8x_{t-1} + \epsilon_t$,

DGP 8 $x_t = x_{t-1} + \epsilon_t$,

DGP 9 $x_t = 0.6\epsilon_{t-1}x_{t-2} + \epsilon_t$,

DGP 10 $x_t = 4x_{t-1}(1 - x_{t-1})$,

DGP 11 $x_t = \sqrt{h_t}\epsilon_t$, $h_t = 1 + 0.8h_{t-1}^2$

m=3	$T = 250$	$T = 500$	$T = 1000$	$T = 1500$	$T = 2000$	$T = 2500$	$T = 3000$
DGP1	0.307	0,598	0,896	0,977	0,995	0,999	1
DGP2	0.138	0.229	0.494	0.709	0,862	0,953	0,98
DGP3	0.054	0,049	0,048	0,048	0,059	0,061	0,066
DGP4	0.717	0,969	1	1	1	1	1
DGP5	0.797	0,981	1	1	1	1	1
DGP6	0.5	0.838	0.988	1	1	1	1
DGP7	0.967	1	1	1	1	1	1
DGP8	0.999	1	1	1	1	1	1
DGP9	0.046	0.046	0.038	0.04	0.049	0.041	0.048
DGP10	1	1	1	1	1	1	1
DGP11	0.077	0.09	0.171	0.204	0.295	0.411	0.447

m=4	$T = 250$	$T = 500$	$T = 1000$	$T = 1500$	$T = 2000$	$T = 2500$	$T = 3000$
DGP1	0.483	0.893	0.998	$0,949$	1	1	1
DGP2	0.213	0.504	0.885	0.993	1	1	1
DGP3	0.114	0.216	$0,514$	$0,809$	$0,949$	$0,985$	$0,997$
DGP4	0.958	1	1	1	1	1	1
DGP5	0.91	0.999	1	1	1	1	1
DGP6	0.7	0.96	1	1	1	1	1
DGP7	0.919	1	1	1	1	1	1
DGP8	1	1	1	1	1	1	1
DGP9	0.226	0.491	0.932	0.996	1	1	1
DGP10	1	1	1	1	1	1	1
DGP11	0.102	0.208	0.409	0.67	0.841	0.926	0.974

m=5	$T = 250$	$T = 500$	$T = 1000$	$T = 1500$	$T = 2000$	$T = 2500$	$T = 3000$
DGP1	0.48	0.924	1	1	1	1	1
DGP2	0.270	0.672	0.985	1	1	1	1
DGP3	0.115	0.323	$0,814$	$0,963$	$0,997$	1	1
DGP4	0.978	1	1	1	1	1	1
DGP5	0.919	1	1	1	1	1	1
DGP6	0.721	0.985	1	1	1	1	1
DGP7	0.999	1	1	1	1	1	1
DGP8	1	1	1	1	1	1	1
DGP9	0.265	0.633	0.971	1	1	1	1
DGP10	1	1	1	1	1	1	1
DGP11	0.102	0.235	0.572	0.83	0.945	0.978	1

References

[1] Denker, M. and Keller, G. (1983) On U-statistics and v. Mises' Statistics for Weakly Dependent Processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 64, 4, pp 505-522.*

[2] Brock, W.A., Dechert, W.D., Scheinkman, J.A. and LeBaron, B. (1996) A test for independence based on the correlation dimension. *Econometric Review, 15, 3, pp 197-235.*

[3] Barnett William A. & Gallant, A. Ronald & Hinich, Melvin J. & Jungeilges, Jochen A. & Kaplan, Daniel T. & Jensen, Mark J., *A single-blind controlled competition among tests for nonlinearity and chaos*, Journal of Econometrics, 82(1),157-192 (1997)

[4] Baxter, M. and Kirchner, T. *Correlation in time-dependent density-functional-theory studies of antiproton-helium collisions*, Phys. Rev. A, 87 (6), 062507 (2013)

[5] Brock W. A., Dechert D., J. A. LeBaron Scheinkman B., Dechert W. D., Scheinkman J. A., *A Test for Independence Based on the Correlation Dimension*, Department of Economics, University of Wisconsin (1996)

[6] De Lima Pedro J. F., *Nuisance Parameter Free properties of Correlation Integral Based Statistics*, Econometric Reviews, 15(3), p. 237-259 (1996)

[7] Figueiredo C., Diambra L., and Pereira C., *Convergence Criterium of Numerical Chaotic Solutions Based on Statistical Measures*, Applied Mathematics, Vol.2, p. 436-443 (2011)

[8] Grassberger Peter and Procaccia Itamar, *Measuring the Strangeness of Strange Attractors*, Physica 9D, pag. 189-208 (1983)

[9] Hiemstra C. and Jones J.D., *Testing for linear and nonlinear Granger causality in the stock price-volumen relation*, Journal of Finance, Vol. 49, p.1639-1665 (1994)

[10] Hsieh, D.A., (1989) *Testing for Nonlinear Dependence in Daily Foreign Exchange Rate Changes*, Journal of Business, 62: 339-368.

[11] Kanzler L., *Very fast and correctly sized estimation of the BDS statistic*, Department of Economics, Oxford Uniersity (1999)

[12] Liu, T., C.W.J. Granger, and W.P. Heller (1998). *Using correlation exponent to decide whether an economic time series is chaotic*, Journal of Applied Econometrics 7: 25-39.

[13] McMillan, D. G. (2003). *Non-linear Predictability of UK Stock Market Returns*, Oxford Bulletin of Economics and Statistics 65(5): 557-573.

[14] Mathew N. and Picu R.C., *Molecular conformational stability in cyclotrimethylene trinitramine crystals*, Journal of Chemical Physics, Vol. 135(2)(2011)

[15] Matilla-Garcia M., Queralt R., Sanz P. and Vazquez F.J., *A Generalized BDS Statistic*, Computational Economics, Vol.24, p.277-300 (2004)

[16] Matilla-Garcia M., *Nonlinear Dynamics in Energy Futures*, The Energy Journal, Vol. 28, No. 3, pag.7-30 (2007)

[17] Matilla-Garcia M., Ruiz Marin, M., Mohammed, D. and Ojeda, RinaB., *Nonparametric correlation integral-based tests for linear and nonlinear stochastic processes*, Decisions in Economics and Finance,1-13. 2013

[18] Tjostheim D., *Measures and tests of independence: a survey*, Statistics, Vol.28, p.249-284 (1996)