The Stata Journal (2021)
21, Number 2, pp. 279–294 DOI: 10.1177/1536867X211025796

The BDS test of independence

Christopher F. Baum
Department of Economics
Boston College
Chestnut Hill, MA
baum@bc.edu

Stan Hurn
School of Economics and Finance
Queensland University of Technology
Brisbane, Australia
s.hurn@qut.edu.au

Kenneth Lindsay
Department of Mathematics
University of Glasgow
Glasgow, UK
kenneth.lindsay@glasgow.ac.uk

Abstract. In this article, we describe and implement the Brock, Dechert, and Scheinkman (1987, Working paper) test of independence of the elements of a time series.

Keywords: st0636, bds, BDS test, dependence, correlation integral

1 Introduction

The question of whether time-series data are independently distributed has a long history in econometrics. In Stata, `runtest` (see [R] **runtest**) is a nonparametric test of the hypothesis that the observations of the series occur in a random order by counting how many successive observations lie above or below a threshold such as the median of the series. Swed and Eisenhart (1943) provide exact critical values for this test.[1] The interest in testing for independence in time series was reinvigorated in the early 1980s by the nonlinear dynamics and chaos literature with the goal of distinguishing deterministic systems from random systems. Brock, Dechert, and Scheinkman (1987) and Brock et al. (1996) introduced a test known as the Brock, Dechert, and Scheinkman (BDS) test for detecting dependence in time-series data based on the correlation dimension of the process. Although the popularity of the BDS test was initially derived from its link with deterministic chaos, it has proved to be useful as a residual diagnostic test because it can detect deviations from dependence in time-series data quite reliably. Furthermore, its asymptotic properties as a residual diagnostic are well understood. Extensive Monte Carlo results have proved the BDS test useful in relatively small samples (Brock, Hsieh, and LeBaron 1991).

[1] A variation on this test analyzes first differences of the series. The "runs-up-and-down" test classifies observations not by whether they lie above or below a threshold but by whether they are steadily increasing or decreasing. Thus, an unbroken string of increases in the variable of interest is counted as one run, as is an unbroken string of decreases. According to Madansky (1988), the runs test is superior to the runs-up-and-down test for detecting trends in the data, but the runs-up-and-down test is superior for detecting autocorrelation. Edgington (1961) has compiled a table of the small-sample distribution of the runs-up-and-down statistic, which is reprinted in Madansky (1988).

The idea behind the BDS is fairly simple. If the data are independently and identically distributed (i.i.d.), then for a given distance ε, the probability that the difference between pairs of m-dimensional points is less than or equal to ε will be a constant. This probability is denoted $C_m(\varepsilon)$ and is known as the correlation integral (Grassberger and Procaccia 1983). The intuition of the BDS test is that with i.i.d. data, $C_m(\varepsilon)$ will simply be the product of the individual pairs, so $C_m(\varepsilon) = C_1(\varepsilon)^m$. However, the sample analogues of these quantities will not satisfy this condition exactly. The BDS test provides a formal statistical evaluation of the significance of this divergence.

Although the intuition of the BDS test is straightforward, calculating the BDS statistic is not easy, and it is even more challenging to perform the computations with the speed necessary to make bootstrap resampling feasible. To complement earlier implementations of the test, LeBaron (1997) describes a fast algorithm and provides C source code to implement the test. The initial component of LeBaron's approach is based on an efficient sorting algorithm that, for a time series of length T, requires $O(T \log T)$ operations and hence pays dividends for large values of T. The main disadvantage of this fast algorithm is that it is not completely transparent.

By contrast, the approach to compute the BDS test described in this article offers transparency of operation because it does not require sorting of the data to estimate the correlation integral. Furthermore, the implementation in Mata makes the procedure accessible on any machine running Stata. This desire to facilitate ease of use does imply a speed penalty. Consequently, our variant is intrinsically $O(T^2)$, which makes it considerably slower than LeBaron's algorithm when applied to a lengthy time series. However, the absence of overhead costs makes the approach efficient for smaller datasets.

In addition to describing the command that implements the test, we also outline an elegant simplification for the computation of the variance of the BDS statistic.

2 The BDS test

The correlation integral introduced by Grassberger and Procaccia (1983) is a method for measuring the frequency with which temporal patterns are repeated in data. Given observations of a time series $x_t,\ t = 1, \ldots T$, the m-history of the time series is

$$x_t^m = (x_t, x_{t+1}, x_{t+2}, \ldots, x_{t+m-1})$$

A time series expressed in this form is described as being embedded of dimension m, with $m = 1$ representing the case where only the original time series is considered. The sample correlation integral for a given $\varepsilon > 0$ and embedding dimension $m = 1$ is

$$\widehat{C}_1 = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} I\big(|x_j - x_k| \leq \varepsilon\big) \tag{1}$$

where $n = (T - m + 1)$ and $I(\mathcal{S})$ is the indicator function taking the value 1 if $\mathcal{S}$ is a true statement and 0 otherwise. For ease of notation, the dependence of this quantity on ε is suppressed. Of course, if ε is chosen so that all pairs satisfy the condition, then

$\widehat{C}_1 = 1$; if ε is chosen so that no pairs satisfy the condition, then $\widehat{C}_1 = 0$. Consequently, the correlation integral has the interpretation of measuring spatial correlation. Often in practice, ε is set in terms of standard deviations of the data.[2]

Strictly speaking, if there are T data points, the computation of $\widehat{C}_1$ with $m = 1$ can use all T elements of the sample to construct the (x_j, x_k) pairs required in (1). However, if higher-order embedding dimensions are to be computed, it is important to ensure that the correlation integrals $\widehat{C}_1, \ldots, \widehat{C}_m$ are each calculated on the same number of (x_j, x_k) pairs. Consequently, the observed sample is always truncated to $x_1, \ldots, x_n$ with $n = (T - M + 1)$, where M is the maximum order of the embedding dimension.

The correlation integral $\widehat{C}_m$ is then given by

$$\widehat{C}_m = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \prod_{r=0}^{m-1} I(|x_{j+r} - x_{k+r}| \leq \varepsilon)$$

Therefore, the mth-order correlation integral computes the joint probability

$$\Pr\big(|x_j - x_k| < \varepsilon, |x_{j+1} - x_{k+1}| < \varepsilon, \ldots, |x_{j+m-1} - x_{k+m-1}| < \varepsilon\big)$$

For a given value of ε, define $I_{j,k} = I(|x_j - x_k| \leq \varepsilon)$. Then, $\widehat{C}_1$ and $\widehat{C}_m$ have simplified expressions

$$\widehat{C}_1 = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} I_{j,k}\,, \qquad \widehat{C}_m = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \prod_{r=0}^{m-1} I_{j+r,k+r} \tag{2}$$

in which occurrences of ε have been suppressed for representational simplicity.

The BDS test proceeds by noting that under the assumption of i.i.d. data, this probability will simply be the product of the individual probabilities for each pair if the observations are independent. Under this null hypothesis, it follows from (3) that

$$\mathrm{E}(\widehat{C}_1) = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \mathrm{E}(I_{j,k}) = \frac{2}{n(n-1)} \sum_{j=1}^{n} \sum_{k=j+1}^{n} C_1 = C_1\,,$$

$$\mathrm{E}(\widehat{C}_m) = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \prod_{r=0}^{m-1} \mathrm{E}(I_{j+r,k+r}) = C_1^m$$

The BDS test provides a formal basis for judging the size of this error. Given a value for ε, Brock et al. (1996) defined the BDS statistic as

$$W_{n,m}(\varepsilon) = \sqrt{n} \left(\frac{\widehat{C}_m - \widehat{C}_1^{\,m}}{\sigma_{n,m}} \right)$$

[2] Sometimes, the data are transformed to the unit interval $[0,1]$ before the test is performed. For each embedding dimension $i = 1 \ldots m-1$, the distance ε is set as 0.9^i, meaning that the test is run over a grid of embedding dimensions and distances; see Cromwell, Labys, and Terraza (1994).

where the standard deviation $\sigma_{n,m}$ is computed from the variance

$$\sigma^2_{n,m} = 4\left\{\beta^m + 2\sum_{j=1}^{m-1}\beta^{m-j}\alpha^{2j} + (m-1)^2\alpha^{2m} - m^2\beta\alpha^{2m-2}\right\} \tag{3}$$

in which α and β are defined as

$$\alpha = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} I_{j,k}\,, \qquad \beta = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} I_{i,j}I_{j,k} \tag{4}$$

Brock et al. (1996) demonstrate that under the null hypothesis of i.i.d. data, this statistic converges in distribution to the standard normal

$$W_{n,m}(\varepsilon) \xrightarrow{d} N(0,1)$$

for a given ε.

3 Simplifying the computation of the variance

It is clear from (3) that the expression for $\sigma^2_{n,m}$ can be regarded as a polynomial of degree m in β with coefficients that are functions of α; that is,

$$\sigma^2_{n,m}(\beta) = 4\left\{\beta^m + 2\sum_{j=1}^{m-1}\beta^{m-j}\alpha^{2j} + (m-1)^2\alpha^{2m} - m^2\beta\alpha^{2m-2}\right\} \tag{5}$$

A straightforward calculation indicates that $\sigma^2_{n,m}(\alpha^2) = 0$. The remainder theorem now asserts that $(\beta - \alpha^2)$ is a factor of (5), which when factored into the expression gives

$$\sigma^2_{n,m}(\beta) = \begin{cases} 4(\beta-\alpha^2)^2 & m = 2 \\ 4(\beta-\alpha^2)g(\beta) & m \geq 3 \end{cases}$$

where $g(\beta)$ is the polynomial of degree $(m-1)$ with expression

$$g(\beta) = \sum_{j=1}^{m-1}(2j-1)\beta^{m-j}\alpha^{2(j-1)} - (m-1)^2\alpha^{2(m-1)}$$

The simplest way to continue the analysis of $\sigma^2_{n,m}(\beta)$ is to replace $(2j-1)$ with the algebraically identical expression $j^2 - (j-1)^2$. With this substitution in place,

$$g(\beta) = \sum_{j=1}^{m-1} j^2\beta^{m-j}\alpha^{2(j-1)} - \sum_{j=1}^{m-1}(j-1)^2\beta^{m-j}\alpha^{2(j-1)} - (m-1)^2\alpha^{2(m-1)}$$

The second summation is reindexed by replacing $j-1$ with j to obtain

$$\begin{aligned} g(\beta) &= \sum_{j=1}^{m-1} j^2\beta^{m-j}\alpha^{2(j-1)} - \sum_{j=1}^{m-2} j^2\beta^{m-j-1}\alpha^{2j} - (m-1)^2\alpha^{2(m-1)} \\ &= \sum_{j=1}^{m-1} j^2\beta^{m-j}\alpha^{2(j-1)} - \sum_{j=1}^{m-1} j^2\beta^{m-j-1}\alpha^{2j} \end{aligned}$$

Both summations are now combined to give

$$\begin{aligned} g(\beta) &= \sum_{j=1}^{m-1} j^2 \left\{\beta^{m-j}\alpha^{2(j-1)} - \beta^{m-j-1}\alpha^{2j}\right\} \\ &= (\beta-\alpha^2)\sum_{j=1}^{m-1} j^2\beta^{m-j-1}\alpha^{2(j-1)} \end{aligned} \tag{6}$$

The case $m=2$ can be incorporated into (6) to derive the final result

$$\sigma^2_{n,m} = 4(\beta-\alpha^2)^2 \sum_{j=1}^{m-1} j^2\beta^{m-j-1}\alpha^{2(j-1)}\,, \quad m \geq 2$$

4 The bds command

The `bds` command calculates the BDS test as described in section 5.

4.1 Syntax

Before using the `bds` command and other similar Stata time-series commands, one must `tsset` or `xtset` the data so that the variable of interest is defined as a proper time series. The command syntax is

`bds` *varname* [*if*] [*in*] [, `m(`*integer*`)` `eps(`*numlist*`)`]

Note that *varname* may not contain gaps within the specified sample. *varname* can contain time-series operators. The command can be applied to one unit of a panel.

4.2 Options

The command supports the following options:

`m(`*integer*`)` specifies the maximum embedding dimension, which evaluates sequences of length 2, . . . , m. The default is `m(3)`.

`eps(`*numlist*`)` provides a list of values of ε to be used in the comparison with the differences of pairs' values. The test evaluates the number of absolute differences that are less than the epsilon value. If *numlist* values are not provided, six defaults are considered: the 70th percentile of the distribution of absolute differences, followed by 0.5, 1.0, 1.5, 2.0, and 2.5 times the standard deviation of the data. If values are provided, they must be given in ascending order.

4.3 Stored results

`bds` stores the following in `r()`:

Scalars	
`r(N)`	number of observations
`r(m)`	maximum embedding dimension
Macros	
`r(cmd)`	`bds`
`r(varname)`	variable name
`r(tsfmt)`	time-series format of the time variable
Matrices	
`r(bds_stat)`	results matrix

5 The algorithm

Given $\varepsilon > 0$, the quantity $I_{j,k}$ has value 1 if $|x_j - x_k| \le \varepsilon$ and 0 otherwise. The first step in the computation of the BDS statistic is to create a lookup table, which in Stata can be visualized as an upper triangular matrix of one-byte integers with T rows and T columns. The (j,k)th entry of this matrix is 1 if $|x_j - x_k| \le \varepsilon$ and 0 otherwise.[3] The (k,k)th diagonal entry of this lookup matrix compares $|x_k - x_k|$ with ε and therefore is 1 for all values of k.

5.1 First-order correlation integral

The sample estimate of the first-order correlation integral is the expected value of $I_{j,k}$ taken over all values of j and k satisfying $j < k \le n$, or equivalently, the fraction of all pairs (x_j, x_k) for which $|x_j - x_k| \le \varepsilon$. Put simply, this is the total count of all the nonzero entries in the upper triangle of the lookup matrix, excluding the main diagonal divided by the total number of distinct pairings, namely, $n(n-1)/2$.

One way to think of this total count is as a sum of row counts $r_1, r_2, \ldots, r_n$, in which r_j denotes the sum of the jth row of the lookup matrix, excluding the contribution from the main diagonal. The alternative view is to think of the total count as a sum of column counts $c_1, c_2, \ldots, c_n$, in which c_j denotes the sum of the jth column of the lookup matrix, excluding the contribution from the main diagonal. In the former, $r_n = 0$, and in the latter, $c_1 = 0$. Both strategies will give the same total count, but each will be composed of different partial counts. The explicit expressions for r_j and c_j are, respectively,

[3] The fast C code provided by LeBaron (1997) stores this table more efficiently as a vector of bits, albeit at the cost of portability and transparency.

$$r_j = \sum_{k=j+1}^{n} I_{j,k}\,, \qquad c_j = \sum_{k=1}^{j-1} I_{k,j}$$

The total number of counts, say, C, is therefore

$$\sum_{j=1}^{n} r_j = C = \sum_{j=1}^{n} c_j \quad \rightarrow \quad 2C = \sum_{j=1}^{n} (r_j + c_j)$$

and the estimated value of the first-order correlation integral is therefore $2C/n(n-1)$.

Value of α

The usual decomposition of a double summation into diagonal and off-diagonal contributions allows α in the first expression in (4) to have representation

$$\alpha = \frac{1}{n^2}\Big(\sum_{j=1}^{n} I_{j,j} + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n} I_{j,k}\Big) = \frac{n+2C}{n^2}$$

Value of β

The calculation of the value of β begins by noting that (4) may be usefully rewritten as

$$\beta = \frac{1}{n^3}\sum_{j=1}^{n}\Big(\sum_{i=1}^{n}\sum_{k=1}^{n} I_{i,j} I_{j,k}\Big)$$

This representation indicates that the value of β is constructed from a count of all triples (x_i, x_j, x_k) for which $I_{i,j} I_{j,k} = 1$, including the possibilities $i = j$, $k = j$, or both. Suppose that this counting process has reached x_j. Then, the previous analysis has demonstrated that there are $m = r_j + c_j + 1$ data points within ε of x_j, including x_j itself. Clearly, the value of m will depend on j, but to maintain representational clarity, we suppress this dependence. Furthermore, suppose that these m data (which include x_j) are $x_{p_1}, \ldots, x_{p_m}$, ignoring again the dependence of $p_1, \ldots, p_m$ on j.

Importantly, $x_{p_1}, \ldots, x_{p_m}$ all share the property that $I_{j,p_1} = I_{p_1,j} \cdots = I_{p_m,j} = I_{j,p_m} = 1$ for a fixed value of j. Consequently,

$$r_j + c_j + 1 = \sum_{k=1}^{m} I_{j,p_k} \quad \rightarrow \quad \Big(\sum_{r=1}^{m} I_{j,p_r}\Big)^2 = (r_j + c_j)^2 + 2(r_j + c_j) + 1$$

However, by construction,

$$\Big(\sum_{r=1}^{m} I_{j,p_r}\Big)^2 = \sum_{r=1}^{m}\sum_{s=1}^{m} I_{p_r,j} I_{j,p_s} = \sum_{i=1}^{n}\sum_{k=1}^{n} I_{i,j} I_{j,k}$$

In conclusion,

$$\beta = \frac{1}{n^3}\sum_{j=1}^{n}\Big(\sum_{i=1}^{n}\sum_{k=1}^{n} I_{i,j}I_{j,k}\Big) = \frac{1}{n^3}\left\{\sum_{j=1}^{n}(r_j+c_j)^2 + 2\sum_{j=1}^{n}(r_j+c_j) + \sum_{j=1}^{n}1\right\}$$
$$= \frac{1}{n^3}\left\{\sum_{j=1}^{n}(r_j+c_j)^2 + 2C + n\right\}$$

5.2 Higher-order correlation integrals

The sample estimate of the correlation integral at embedding dimension $m \le M$ is the expected value of all quantities of type

$$\prod_{p=0}^{m-1} I_{j+p,k+p} \tag{7}$$

taken over the $n(n-1)/2$ possible values of (j,k) for which $1 \le j < k \le n$, or equivalently, the fraction of all pairs (x_j, x_k) for which expression (7) is 1, that is, $I_{j+p,k+p} = 1$ for all values of p satisfying $0 \le p < m$. Given a pair (x_j, x_k), the contributions made by that pair to the correlation integrals at each depth of embedding may be calculated simultaneously by noting that

$$\prod_{p=0}^{m-1} I_{j+p,k+p} = \Big(\prod_{p=0}^{m-2} I_{j+p,k+p}\Big) \times I_{j+m-1,k+m-1} \tag{8}$$

This means that the computational penalty involved in the calculation of higher-order correlation integrals is small, provided the contributions from each pair, say, (x_j, x_k), at all requested levels of embedding are done simultaneously.

The matrix representation of the lookup table allows (8) to be computed efficiently. Because $(k+p)-(j+p) = k-j = r$ is independent of p (and consequently the depth of embedding), the value of $I_{j+p,k+p}$ is the entry in the matrix lookup table at row $(j+p)$ and column $(j+p)+r$. In overview, the contributions to the estimated correlation integral at embedding depth m can be visualized as the sum of the products of m consecutive elements of the super diagonals of the matrix lookup table taken over the entire upper triangle (that is, excluding the main diagonal) for values of m from $m=2$ to $m=M$. The estimate of the correlation integral is this sum divided by $n(n-1)/2$.

6 Empirical applications

6.1 Sunspots

Sunspots are regions on the surface of the sun with magnetic field strengths thousands of times stronger than the earth's magnetic field. They appear as dark spots on the surface of the sun and typically last for several days. The data, y_t, which are the annual averages of daily sunspot numbers from 1700 to 2017, were compiled by the Solar Influences Data Analysis Center in Belgium. The series is plotted in figure 1. This series not only is the longest directly observed index of solar activity but also has interesting time-series properties.

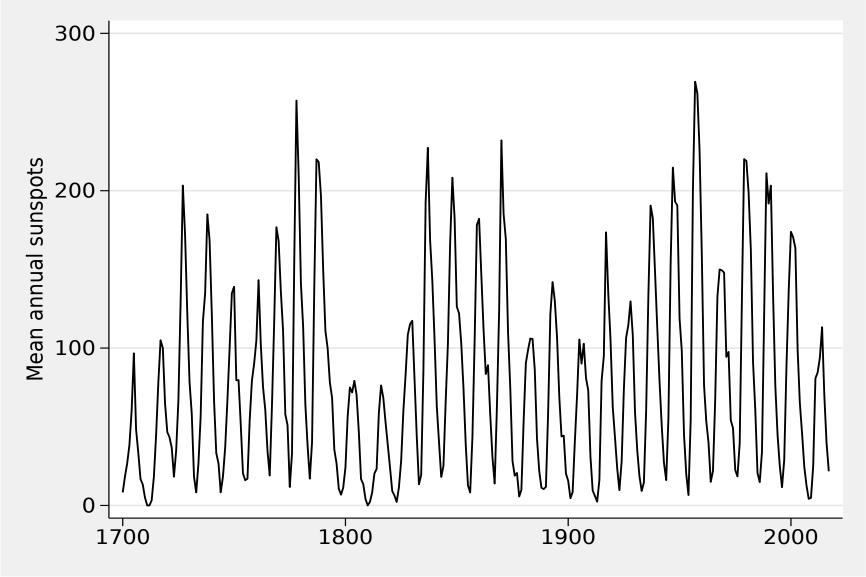

Figure 1. Plot of average annual sunspots series from 1700 to 2017

The sunspot numbers have been of interest to climatologists who hypothesized a link between sunspot activity and climate change. White and Liu (2008) provide evidence that the solar cycle may be the trigger for El Niño and La Niña episodes from 1900–2005, suggesting that higher solar activity implies weaker and less frequent El Niño events. This view is controversial, and no generally accepted statistical link has been established. The so-called Maunder Minimum[4] between 1645 and 1715 was a period in which sunspots were scarce and the winters harsh, strongly suggesting a link between solar activity and climate change. The current view is that there has been no significant long-term upward trend in solar activity since 1700. This implies that rising global temperatures since the industrial revolution cannot be attributed to increased solar activity.

A simple application of the BDS test to the annual sunspot data confirms the strong rejection of the null hypothesis of i.i.d. data. The value of ε is set to one standard deviation of the data, and the maximum embedding dimension is 4.

[4] The Maunder Minimum is named after the solar astronomers Annie Russell Maunder (1868–1947) and her husband, Edward Walter Maunder (1851–1928), who studied how sunspot latitudes changed with time.

```
. bds sunspots, eps(61.985) m(4)
Brock, Dechert, Scheinkman test for independence
N(0,1] test statistics for sunspots, n (adjusted) = 315, sd = 61.98554

               eps    m     BDSstat     stderr  z-value        count
---------+------------------------------------------------------------
61.985   |    61.98   2     33.7834     0.0032   0.0000        19509
       _ |    61.98   3     35.6796     0.0039   0.0000        14438
       _ |    61.98   4     40.7236     0.0036   0.0000        11167
```

In fact, the sunspot data have given rise to many attempts to use nonlinear modeling to capture the key features of the time series. In particular, many different self-exciting threshold autoregressive models have been proposed in the literature and applied to the sunspot data. Consider the following model, which is similar to those of Tong and Lim (1980) and Battaglia and Orfei (2005):

$$y_t = \begin{cases} \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + v_{1t} & \text{if } y_{t-3} \leq k \\ \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + v_{2t} & \text{if } y_{t-3} > k \end{cases} \tag{9}$$

Fitting the model using the Stata command `threshold` (see [TS] **threshold**) yields

```
. threshold sunspots, threshvar(L3.sunspots) regionvars(L(1/2).sunspots) nodots
> cformat(%8.5f)
Searching for threshold: 1
(Running 252 regressions)
Threshold regression
                                                 Number of obs     =        315
Full sample:    1703 - 2017                      AIC               =  1981.3498
Number of thresholds =  1                        BIC               =  2003.8652
Threshold variable: L3.sunspots                  HQIC              =  1990.3456

-----------------------------------
Order     Threshold         SSR
-----------------------------------
1         60.7000       1.635e+05
-----------------------------------

------------------------------------------------------------------------------
    sunspots |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Region1      |
    sunspots |
         L1. |    1.57704    0.06118    25.78   0.000      1.45713     1.69694
         L2. |   -1.05771    0.09999   -10.58   0.000     -1.25369    -0.86174
             |
       _cons |   32.28765    2.75804    11.71   0.000     26.88200    37.69330
-------------+----------------------------------------------------------------
Region2      |
    sunspots |
         L1. |    1.02593    0.06255    16.40   0.000      0.90333     1.14853
         L2. |   -0.20113    0.06635    -3.03   0.002     -0.33116    -0.07109
             |
       _cons |   -3.10774    4.24440    -0.73   0.464    -11.42662     5.21114
------------------------------------------------------------------------------
```

The optimal threshold value returned by the search is $k = 60.7$, which is slightly larger than the values reported by Tong and Lim (1980) and Battaglia and Orfei (2005), although they use annual data over a shorter sample period in their work and include a much more complex dynamic structure. The autoregressive coefficients in both regimes have a somewhat similar pattern in terms of sign, although the difference in their sizes is slightly more marked in regime 2. These results suggest a rather complex dynamic pattern, although it is possible to conjecture that the large positive first-order autocorrelation coefficient is offset by a much smaller negative second-order autocorrelation coefficient, which implies that the process persists a little longer in regime 2.

The residuals from the threshold regression are plotted in figure 2. The first impression is that there is less structure in the residuals than in the original series, but the real question to be addressed is whether the residuals are now i.i.d. Application of the BDS, again with a maximum embedding dimension of 4 specified as one of the options, yields the following results:

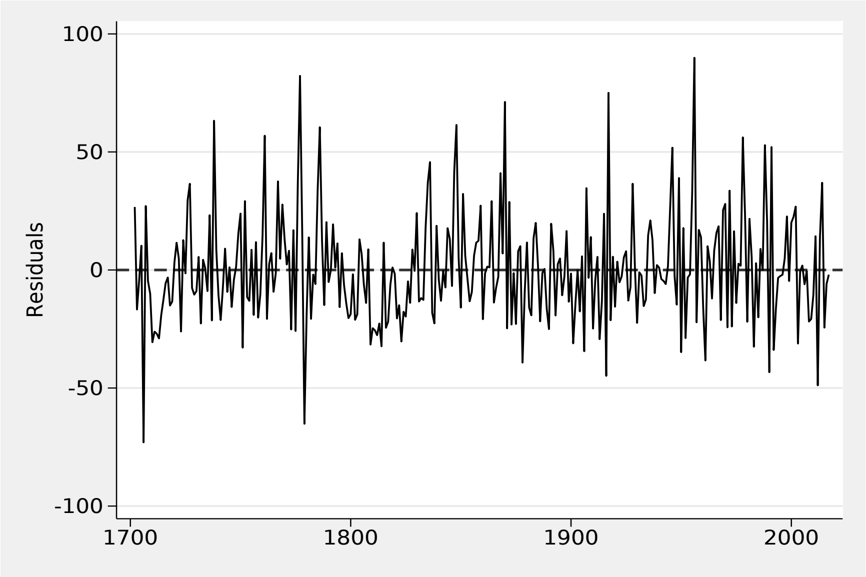

Figure 2. Residuals obtained from fitting the model in (9) using the annual sunspot data

```
. bds uhat, m(4)
Brock, Dechert, Scheinkman test for independence
N(0,1] test statistics for uhat, n (adjusted) = 313, sd = 22.83103
               eps   m     BDSstat     stderr  z-value         count
─────────┼──────────────────────────────────────────────────────────
  fp0.7       30.84  2      3.7536     0.0045   0.0002         24501
      _       30.84  3      3.7239     0.0071   0.0002         17775
      _       30.84  4      3.6170     0.0084   0.0003         12955
  0.5sd       11.42  2      6.0857     0.0016   0.0000          4960
      _       11.42  3      6.3498     0.0011   0.0000          1706
      _       11.42  4      6.5673     0.0006   0.0000           599
  1.0sd       22.83  2      4.3457     0.0041   0.0000         16173
      _       22.83  3      4.1801     0.0053   0.0000          9633
      _       22.83  4      4.0002     0.0050   0.0001          5768
  1.5sd       34.25  2      3.4208     0.0044   0.0006         27676
      _       34.25  3      3.5086     0.0073   0.0005         21271
      _       34.25  4      3.4108     0.0092   0.0006         16399
  2.0sd       45.66  2      2.1676     0.0033   0.0302         35987
      _       45.66  3      2.4953     0.0063   0.0126         31214
      _       45.66  4      2.4210     0.0090   0.0155         27080
  2.5sd       57.08  2      0.5118     0.0020   0.6088         41227
      _       57.08  3      0.9191     0.0042   0.3580         38001
      _       57.08  4      0.8492     0.0065   0.3958         34992
```

For choices of ε up to two standard deviations, the test statistic is significant and the null hypothesis of i.i.d. is strongly rejected. This result is a warning that the apparent lack of structure obtained from a visual impression can be misleading. Interestingly, at 2.5 standard deviations, the null hypothesis cannot be rejected. It may be that this choice of ε is simply too large for these data.

6.2 U.S. equity returns

Consider the example given in Hurn et al. (2020), in which real U.S. equity returns are to be forecast using an autoregressive [AR(1)] model based on the assumption of normality. The model is

$$r_t = \phi_0 + \phi_1 r_{t-1} + v_t\,, \qquad v_t \sim N(0, \sigma_v^2)$$

If the observed values r_t are indeed generated correctly according to this simple model, then the transformed quantity

$$u_t = \Phi\left(\frac{v_t}{\sigma_v}\right), \qquad t = 1, 2, \dots, T$$

takes values in the unit interval $[0, 1]$ because of the fundamental property of $\Phi(\cdot)$, which is the cumulative distribution function (CDF) of the standard normal distribution. This transformation is known as the probability integral transform (Diebold, Gunther, and Tay 1998). Furthermore, Rosenblatt (1952) demonstrates that if the model is correctly specified, the probability integral transform, u_t, will be independent and uniformly distributed on the unit interval. The null hypothesis that the model is correctly specified can then be tested in terms of the BDS test applied to u_t.

Using monthly observations for the period January 1871 to September 2016 on real equity returns, we find estimation of the model gives the following results.[5]

```
. regress re L1.re, cformat(%8.5f)
      Source │       SS           df       MS      Number of obs   =     1,747
─────────────┼──────────────────────────────────   F(1, 1745)      =    153.00
       Model │  2331.34477         1  2331.34477   Prob > F        =    0.0000
    Residual │  26590.2296     1,745  15.2379539   R-squared       =    0.0806
─────────────┼──────────────────────────────────   Adj R-squared   =    0.0801
       Total │  28921.5744     1,746  16.5644756   Root MSE        =    3.9036

─────────────┬────────────────────────────────────────────────────────────────
          re │      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
─────────────┼────────────────────────────────────────────────────────────────
          re │
         L1. │    0.28392    0.02295    12.37   0.000      0.23890     0.32894
             │
       _cons │    0.25270    0.09375     2.70   0.007      0.06883     0.43657
─────────────┴────────────────────────────────────────────────────────────────
```

The probability integral transform is given by

$$u_t = \Phi\left(\frac{\widehat{v}_t}{\widehat{\sigma}_v}\right)$$

in which $\widehat{\sigma}_v$ is the standard error of the regression. The BDS for independence applied to u_t with embedding dimension $m = 4$ gives

```
. bds u1, m(4)
Brock, Dechert, Scheinkman test for independence
N(0,1] test statistics for u1, n (adjusted) = 1744, sd = .2568086
          │      eps   m     BDSstat      stderr  z-value          count
──────────┼──────────────────────────────────────────────────────────────
  fp0.7   │     0.39   2      5.2523      0.0011   0.0000         752008
       _  │     0.39   3      6.8475      0.0017   0.0000         537662
       _  │     0.39   4      7.9217      0.0020   0.0000         387828
  0.5sd   │     0.13   2      7.6641      0.0002   0.0000         108847
       _  │     0.13   3     10.8246      0.0001   0.0000          30401
       _  │     0.13   4     13.3189      0.0001   0.0000           8717
  1.0sd   │     0.26   2      6.5204      0.0008   0.0000         387647
       _  │     0.26   3      8.4072      0.0009   0.0000         201017
       _  │     0.26   4      9.7706      0.0007   0.0000         105768
  1.5sd   │     0.39   2      5.2760      0.0011   0.0000         733400
       _  │     0.39   3      6.8436      0.0017   0.0000         517956
       _  │     0.39   4      7.9101      0.0020   0.0000         369116
  2.0sd   │     0.51   2      5.0026      0.0009   0.0000        1055275
       _  │     0.51   3      6.6070      0.0017   0.0000         887570
       _  │     0.51   4      7.6394      0.0023   0.0000         750281
  2.5sd   │     0.64   2      5.0064      0.0004   0.0000        1296845
       _  │     0.64   3      6.7856      0.0009   0.0000        1202422
       _  │     0.64   4      7.9056      0.0013   0.0000        1117164
```

The null hypothesis is rejected, and the conclusion is that the AR(1) model of equity returns is misspecified because u_t is not i.i.d. A histogram of the transformed time series, u_t, given in figure 3 suggests that the distribution of u_t is also not uniform. The interior

[5] The data are obtained from the website of Robert J. Shiller at http://www.econ.yale.edu/~shiller/.

peak of the distribution of u_t and also the peak at zero suggest that equity returns, r_t, are not consistent with the specification of an AR(1) model with normally distributed errors.

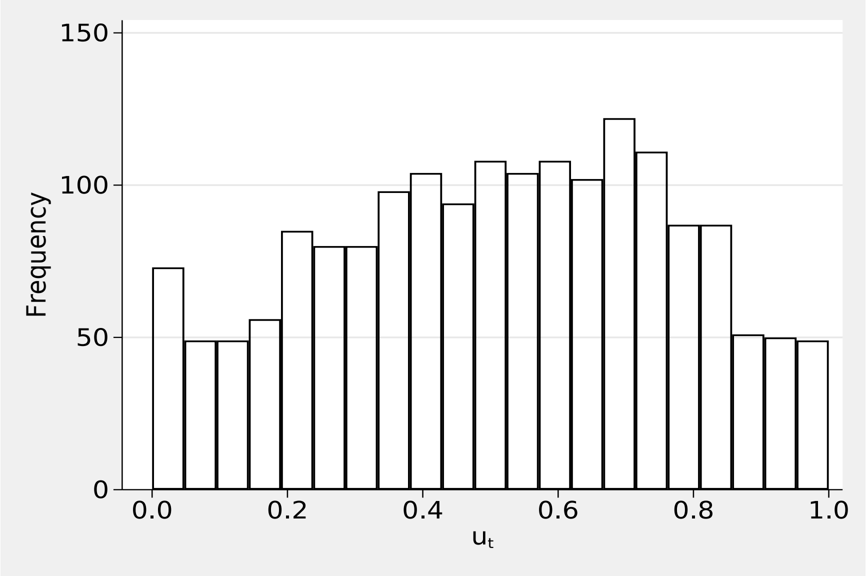

Figure 3. Probability integral transform applied to the forecast errors of the AR(1) model of United States equity returns, January 1871 to September 2016

An interesting feature of the results is that the BDS statistic is much smaller for the choice of ε based on the fraction of pairs rather than on the standard deviations. This may be suggestive that the fraction of pairs method for choosing ε is more robust to the distribution of the underlying series.

7 Conclusion

In this article, we introduced the `bds` command, which computes the BDS test of the null hypothesis that the time series to which it is applied consists of i.i.d. observations. The algorithm is adapted from LeBaron's C codes and produces the same output as these codes. In addition to providing the command, we also provided an elegant simplification of the computation of the variance of the statistic.

The fact that the procedure does not use compiled code but is written in Mata to run on all machines on which Stata is loaded means that the routine is computationally less efficient when the sample size is large. Consequently, this routine is not suitable for large-scale simulation studies. Two empirical examples demonstrated how the routine is implemented in practice.

8 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 21-2
. net install st0636      (to install program files, if available)
. net get st0636          (to install ancillary files, if available)
```

9 References

Battaglia, F., and L. Orfei. 2005. Outlier detection and estimation in nonlinear time series. *Journal of Time Series Analysis* 26: 107–121. https://doi.org/10.1111/j.1467-9892.2005.00392.x.

Brock, W. A., W. D. Dechert, and J. A. Scheinkman. 1987. A test for independence based on the correlation dimension. Working paper, Department of Economics, University of Wisconsin–Madison.

Brock, W. A., W. D. Dechert, J. A. Scheinkman, and B. LeBaron. 1996. A test for independence based on the correlation dimension. *Econometric Reviews* 15: 197–235. https://doi.org/10.1080/07474939608800353.

Brock, W. A., D. A. Hsieh, and B. LeBaron. 1991. *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence*. Cambridge, MA: MIT Press.

Cromwell, J. B., W. C. Labys, and M. Terraza. 1994. *Univariate Tests for Time Series Models*. Quantitative Applications in the Social Sciences. Thousand Oaks, CA: SAGE. https://dx.doi.org/10.4135/9781412986458.

Diebold, F. X., T. A. Gunther, and A. S. Tay. 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39: 863–883. https://doi.org/10.2307/2527342.

Edgington, E. S. 1961. Probability table for number of runs of signs of first differences in ordered series. *Journal of the American Statistical Association* 56: 156–159. https://doi.org/10.1080/01621459.1961.10482102.

Grassberger, P., and I. Procaccia. 1983. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena* 9: 189–208. https://doi.org/10.1016/0167-2789(83)90298-1.

Hurn, S., V. Martin, P. C. B. Phillips, and J. Yu. 2020. *Financial Econometric Modeling*. New York: Oxford University Press.

LeBaron, B. 1997. A fast algorithm for the BDS statistic. *Studies in Nonlinear Dynamics and Econometrics* 2: 53–59. https://doi.org/10.2202/1558-3708.1029.

Madansky, A. 1988. *Prescriptions for Working Statisticians*. New York: Springer.

Rosenblatt, M. 1952. Remarks on a multivariate transformation. *Annals of Mathematical Statistics* 23: 470–472. https://doi.org/10.1214/aoms/1177729394.

Swed, F. S., and C. Eisenhart. 1943. Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics* 14: 66–87. https://doi.org/10.1214/aoms/1177731494.

Tong, H., and K. S. Lim. 1980. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society,* Series B 42: 245–268. https://doi.org/10.1142/9789812836281_0002.

White, W. B., and Z. Liu. 2008. Non-linear alignment of El Niño to the 11-yr solar cycle. *Geophysical Research Letters* 35. https://doi.org/10.1029/2008GL034831.

About the authors

Christopher F. Baum is a professor of economics and social work at Boston College and DIW research fellow at the German Institute for Economic Research. He is the author of two Stata Press books and maintains the Statistical Software Components archive of community-contributed software.

Stan Hurn is a professor of econometrics in the School of Economics and Finance at Queensland University of Technology and the director of the National Centre for Econometric Research in Australia.

Kenneth Lindsay is an honorary senior research fellow in the School of Mathematics and Statistics at the University of Glasgow.