

Supplement to Accompany:

**Complexity and the Character of Stock Returns: Empirical Evidence and
A Model of Asset Prices Based Upon Complex Investor Learning
Revised 10/11/05**

Scott C. Linn*

Michael F. Price College of Business, University of Oklahoma, Norman, OK 73019 USA

Nicholas S. P. Tay

School of Business and Management, University of San Francisco, CA 94117 USA

*Corresponding author:

Professor Scott C. Linn
Division of Finance
Michael F. Price College of Business
205A Adams Hall
University of Oklahoma
Norman, OK 73019
USA

Tel.: 1 405 325 3444/ 1 405 325 5591; fax: 1 405 325 2096; e-mail: slinn@ou.edu

Contents

Appendix A

The Correlation Integral and the BDS Test Statistic

Appendix B

Construction of the Jarque-Bera statistic, Ljung-Box Q statistic, the ARCH LM statistic and the Long Memory Test statistic V of Lo (1991)

Appendix C

Implementation of Combination Experiments and Individual Hypothesis Experiments and Stochastic Universal Sampling

Appendix A

The Correlation Integral and the BDS Test Statistic

The BDS test (Brock et al. (1987), Brock et al. (1996)) is a test for dependence through time in a series. We begin by choosing what we call a distance parameter, ε . Next consider a pair of consecutive points in the series. If the observations of the series truly are IID, then for any pair of consecutive points, the probability of the distance between these points being less than or equal to ε will be constant. Denote this probability by $C_1(\varepsilon, T)$ for a series of length T , x_t , with $t=1, \dots, T$.

Grassberger and Procaccia (1983a, 1983b) have demonstrated that one may “embed” or reconstruct the trajectory of x_t in an n -dimensional “embedding space” by organizing the observations into n -histories x_t^n , defined by $x_t^n = (x_t, \dots, x_{t+n-1})$. A series of length T contains at most $\text{int}(T/n)$ non-overlapping points where $\text{int}(\bullet)$ defines the maximum integer associated with the calculation T/n . Consider a comparison of the points in two trajectories, one beginning at date t and one beginning at date $q = t+l$. Define the set of matched points as

$$(x_t, x_q), (x_{t+1}, x_{q+1}), \dots, (x_{t+n-1}, x_{q+n-1}). \tag{A.1}$$

Define the joint probability of every pair of points satisfying the ε distance condition by the probability $C_n(\varepsilon, T)$. If the data are IID then $C_1(\varepsilon, T)^n = C_n(\varepsilon, T)$.

The correlation integral is a measure of the fraction of the points in the n -dimensional embedding space with values within the distance ε of each other. A series that is described by a nonlinear generating process is likely to exhibit a greater incidence of clustering in the data than will a linear generating process (Brock, Hsieh and LeBaron (1991), Campbell, Lo and MacKinlay (1997)). Grassberger and Procaccia propose calculating the correlation integral, $C_n(\varepsilon)$, as follows.

$$C_n(\varepsilon) = \lim_{T \rightarrow \infty} \frac{2}{T_n(T_n - 1)} \sum_{i < j} I(x_i^n, x_j^n, \varepsilon) \quad (\text{A.2})$$

where $T_n = T - (n - 1)$, and $I(x_i^n, x_j^n, \varepsilon)$ is an indicator function that equals one if for a pair of n -histories x_i^n and x_j^n , commencing at date i (j) respectively, the greatest absolute difference,

$$\max_{p=0, \dots, n-1} |x_{i-p} - x_{j-p}|, \text{ between the corresponding members of the pair is smaller than } \varepsilon.$$

The BDS test takes as the null hypothesis that the data are independent and identically distributed. Brock, Dechert and Scheinkman (1987) and Brock, Dechert, LeBaron and Scheinkman (1996) show that under the null hypothesis the correlation integral obeys the following relation

$$C_n(\varepsilon, T) \rightarrow C_1(\varepsilon, T)^n \text{ with probability one as } T \rightarrow \infty \quad (\text{A.3})$$

for fixed values of n (the embedding dimension) and ε where $C_n(\varepsilon, T)$ is the correlation integral. Brock et al (1996) go on to show that under the null hypothesis of IID data the statistic

$$BDS_n(\varepsilon, T) = \sqrt{T} \left[C_n(\varepsilon, T) - C_1(\varepsilon, T)^n \right] / \sigma_n(\varepsilon, T) \quad (\text{A.4})$$

has a standard normal distribution in the limit. Consistent estimates of the parameters of the BDS statistic can be computed from the data (Brock, Dechert and Scheinkman (1987), Hsieh (1991), Brock, Dechert, LeBaron and Scheinkman (1996), LeBaron (1997)).

References

- Brock, W. A.; W. D. Dechert; and J. Scheinkman. "A Test for Independence Based upon the Correlation Dimension." Working Paper, University of Wisconsin (1987).
- Brock, W. A.; W. D. Dechert; B. LeBaron; and J. Scheinkman. "A Test for Independence Based upon the Correlation Dimension." *Econometric Reviews*, 15 (1996), 197-235.
- Hsieh, D. A. "Chaos and Nonlinear Dynamics: Application to Financial Markets." *Journal of Finance*, 46 (1991), 1839-1877.

LeBaron, B. "A Fast Algorithm for the BDS Statistic." *Studies in Nonlinear Dynamics and Econometrics*, 2 (1997), 53-59.

Appendix B

Construction of the Jarque-Bera statistic, the Ljung-Box Q statistic, the ARCH LM statistic and the Long Memory Test statistic V of Lo (1991)

The Jarque Bera statistic

The Jarque-Bera statistic is used to test the hypothesis that a given set of data is drawn from a normal distribution. The test statistic measures the difference of the skewness and kurtosis of the series with those from the normal distribution. The statistic is computed as:

$$JB = \frac{N-k}{6} \left(S^2 + \frac{1}{4}(K-3)^2 \right)$$

where S is the skewness, K is the kurtosis, and k represents the number of estimated parameters used to create the series (Jarque and Bera (1987)). Under the null hypothesis of a normal distribution, the Jarque-Bera statistic is distributed as χ^2 with 2 degrees of freedom.

The Ljung-Box Q statistic

The statistic ρ_ℓ refers to the autocorrelation of the relevant series at lag ℓ . $Q(q)$ refers to the Q -statistic computed up to a lag q . The Q -statistic is computed in the following manner (Ljung and Box (1979)):

$$Q_q = T(T-2) \sum_{\ell=1}^q \frac{\rho_\ell^2}{(T-\ell)}$$

Under the null hypothesis of no autocorrelation the Q -statistic is distributed χ^2 with q df.

Let \hat{e}_t be the computed error from a model fitted to the data.

The ARCH LM statistic of Engle

The ARCH LM test of Engle (1982) equals TR^2 where T is the length of the time series and R^2 is obtained from a regression of $\hat{\epsilon}_t^2$ on its q lagged values. Both statistics are distributed as χ_q^2 where the degrees of freedom q equals the number of lags used in computing the statistic.

The V statistic of Lo (1991)

Let r_t , $t = 1, \dots, n$, represent a series of asset returns with mean $\bar{r} = (1/n) \sum_t r_t$. Lo's modified rescaled range statistic equals

$$V_{n,q} = \left(\frac{1}{\hat{\sigma}_n(q)} \right) \left[\text{Max} \sum_{j=1}^k (r_j - \bar{r}_n) - \text{Min} \sum_{j=1}^k (r_j - \bar{r}_n) \right]$$

(4.1)

for $1 \leq i \leq n$ where

$$\hat{\sigma}_n^2(q) = \frac{1}{n} \sum_{j=1}^n (r_j - \bar{r}_n)^2 + \frac{2}{n} \sum_{j=1}^q \omega_j(q) \left[\sum_{i=j+1}^n (r_i - \bar{r}_n)(r_{i-j} - \bar{r}_n) \right] =$$

$$\hat{\sigma}_n^2(q) = \hat{\sigma}_r^2 + 2 \sum_{j=1}^q \omega_j(q) \hat{\gamma}_j$$

and $\omega_j(q) = 1 - j/(q+1)$ for $q < n$, $\hat{\sigma}_r^2$ and $\hat{\gamma}_j$ are the sample variance and lag j autocovariance. The estimator $\hat{\sigma}_n^2(q)$ involves both the variance as well as a weighted sum of the autocovariances where the weights are those suggested by Newey and West (1987) for optimal estimation of the variance of a series in the presence of autocorrelation.

References

- Engle, R. F. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica*, 50 (1982), 987-1007.
- Jarque, C. M., and A. K. Bera. "A Test for Normality of Observations and Regression Residuals." *International Statistical Review*, 55 (1987), 163-172.
- Ljung, G., and G. Box. "On A Measure of Lack of Fit in Time Series Models." *Biometrika*, 66 (1979), 265-270.
- Lo, A. "Long-Term Memory in Stock Market Prices." *Econometrica*, 59 (1991), 1279-1313.

Appendix C

Implementation of Combination Experiments and Individual Hypothesis Experiments and Stochastic Universal Sampling

Stochastic Selection Algorithms

The following discussion on stochastic selection borrows heavily from *GEATbx: Genetic and Evolutionary Algorithm Toolbox for use with MATLAB Version 3.50* (July 2004), Hartmut Pohlheim (<http://www.geatbx.com/docu/index.html>).

Selection of hypotheses for combination or for individual perturbation in the experiments we conduct is based upon forecast accuracy. Each hypothesis in the selection pool receives a probability of selection that depends on its own accuracy as well as the accuracy values of the agent's other hypotheses.

Roulette-Wheel Selection

The simplest selection scheme is roulette-wheel selection, also called stochastic sampling with replacement, Baker (1987). Roulette-Wheel selection is a stochastic algorithm that involves the following process: The hypotheses are mapped to contiguous segments of a line, such that each hypothesis's segment is equal in size to its accuracy value. A random number is generated and the hypothesis whose segment spans the random number is selected. The process is repeated until the number of hypotheses desired is selected. This technique is analogous to a roulette wheel with each slice of the wheel proportional in size to the accuracy of each hypothesis.

Table A shows hypothetical accuracy values and selection probabilities for 11 hypotheses. Hypothesis 1 is the most accurate and occupies the largest interval whereas hypothesis 10 as the second least accurate has the smallest interval on the line (see Figure A). Hypothesis 11, the least accurate, has an accuracy value of 0 and hence has no chance of selection.

Table A Hypothetical Accuracy Values and Selection Probabilities

Number of hypotheses	1	2	3	4	5	6	7	8	9	10	11
Accuracy value	2.0	1.8	1.6	1.4	1.2	1.0	0.8	0.6	0.4	0.2	0.0
Selection probability	0.18	0.16	0.15	0.13	0.11	0.09	0.07	0.06	0.03	0.02	0.0

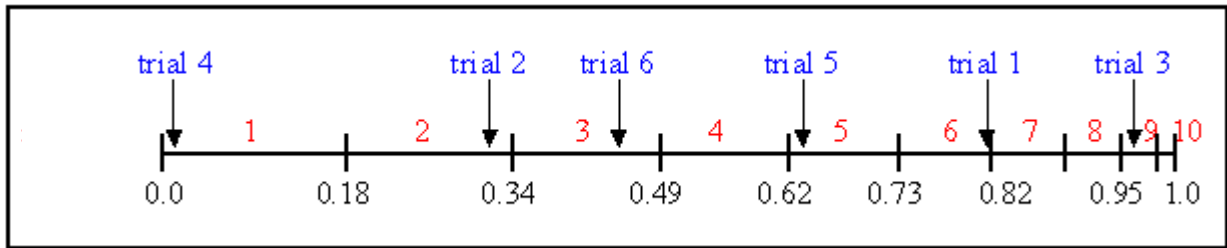
The Selection Probabilities are the individual fitness values normalized by the sum of the fitness values for the 11 hypotheses. Construct a line segment with discrete points showing the cumulative selection probability (see lower half of Figure A).

Suppose we wish to select 6 of the 11 hypotheses. Selection occurs by drawing 6 numbers from a uniform distribution $U[0, 1]$.

Consider a sample of 6 pseudo-random numbers drawn from $U[0,1]$: **0.81, 0.32, 0.96, 0.01, 0.65, 0.42**. Map these points onto the line showing the cumulative selection probabilities.

Figure A shows the selection process of the individuals for the example in Table A together with the above sample trials.

Fig. A Roulette-Wheel Selection



Integers represent hypotheses. Source: Pohlheim (2004).

The hypotheses selected are: 1, 2, 3, 5, 6, 9.

The roulette-wheel selection algorithm provides a zero bias in the sense that the probability of selection is based upon relative accuracy size so that over a large number of trials, say Q , each hypothesis will on average be represented in the same proportion as its ‘Selection probability’ (see Table A). Because the selection process is stochastic in nature, in any given trial the proportional representation of any hypothesis may however deviate from the ‘Selection probability’. Suppose the proportions represented by a hypothesis i over Q trials are given by the vector $pr_i = [pr_{i,1} \ pr_{i,2} \ pr_{i,3} \ \dots \ pr_{i,Q}]$. Define the variance of pr_i as the ‘spread’. The roulette-wheel selection algorithm does not insure minimum ‘spread’, Baker (1987).

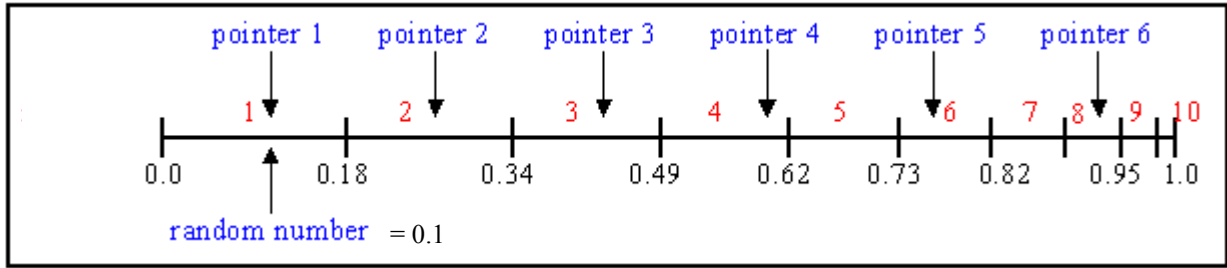
Stochastic Universal Sampling (SUS)

Stochastic universal sampling, Baker (1987), provides zero bias and minimum spread. The hypotheses are mapped to contiguous segments of a line, such that each hypothesis’s segment is equal in size to its accuracy exactly as in roulette-wheel selection. Here however equally spaced pointers are placed over the line, as many as there are objects to be selected. Let N be the number of hypotheses to be selected. The distance between the pointers is equal to $1/N$. The position of the first pointer is given by a random drawing from $U[0, 1/N]$. For $N=6$, the distance between the pointers is therefore $1/6=0.167$. Figure B shows the selection process for the above example.

Begin by drawing a single random number from $U [0, 0.167]$. Suppose this draw equals **0.1**.

Next compute the pointers, where each pointer is $1/N = 0.167$ distance from the previous pointer, and the first pointer is located at the point defined by the above drawing, **0.1**. Map the pointers onto the line defined by the cumulative sums of the selection probabilities.

Fig. B Stochastic Universal Sampling



Integers represent hypotheses. Source: Pohlheim (2004).

The selected hypotheses are therefore: 1, 2, 3, 4, 6, 8.

Sampling in the Experiments

In our experiments a random selection of N objects from a population of size T is desired. Stochastic Universal Sampling is employed. Define a $1 \times T$ vector where each cell holds an identifier for each object from the most-fit object to the least fit object. We label this vector T . Let the identifier for object i be defined by d_i , and define $S_\tau = \sum_{i=1}^{\tau} d_i / \sum_{i=1}^T d_i$ so that the vector $S = [S_1 \ S_2 \ S_3 \ \dots \ 1]$ contains the cumulations of the normalized fitness values for the T objects (see Fig. B above for illustration). The SUS method selects the desired N objects of interest in one step. The one-shot draw is implemented by the use of N equally spaced pointers separated by the distance $1/N$, where N is the desired number of objects. As in the illustration above a number \hat{t} is drawn from $U[0, 1/T]$ followed by the computation of the N equally spaced pointers which are placed in the vector

$$\hat{p} = \left[\hat{t} \quad \hat{t} + \frac{1}{N} \quad \hat{t} + \frac{2}{N} \quad \hat{t} + \frac{3}{N} \quad \dots \quad \hat{t} + \frac{N-1}{N} \right].$$

The N desired objects from the total T objects are then chosen by whether the pointers in \hat{p} lie within the intervals defined by the points in the vector S (see Fig. B above for illustration).

Combination Experiments

Combination Experiments occur with a probability of $\text{Pr}(\text{Comb})$ for any given agent at each date. If an experiment is to take place for an agent, we do the following: 1) select two hypotheses out of the 5 hypotheses held by the agent; selection is performed with the SUS selection process described above based on the fitness values of the agent's hypotheses. Once the two hypotheses are selected uniform crossover is conducted. Uniform crossover is implemented by first creating a **Mask** that is a row vector of size 1×7 . Each cell is randomly filled with a 0 or 1. This is implemented by drawing a vector (1×7) of random numbers from $U[0,1]$ and setting those numbers that are greater than 0.5 to ones and the remaining to zeros. The vector is of length 7 because we have 5 conditions and 2 consequences for each rule of a hypothesis. The cells of the vector that contain 0's are then filled with the information contained in the matching positions of the vectors for each rule of the first selected hypothesis, and the cells containing 1's are filled with the information contained in the matching positions from the second selected hypothesis. An

example will help illustrate the process. Suppose that Hypotheses 1 and 2 are selected for combination, where

$$\text{Hypothesis 1:} \begin{array}{l} \text{Rule 1} \\ \text{Rule 2} \\ \text{Rule 3} \\ \text{Rule 4} \end{array} \begin{bmatrix} 1 & 3 & 0 & 2 & 4 & 1 & 3 \\ 3 & 2 & 0 & 1 & 3 & 2 & 1 \\ 4 & 4 & 0 & 3 & 1 & 4 & 2 \\ 2 & 1 & 0 & 4 & 2 & 3 & 4 \end{bmatrix}$$

$$\text{Hypothesis 2:} \begin{array}{l} \text{Rule 1} \\ \text{Rule 2} \\ \text{Rule 3} \\ \text{Rule 4} \end{array} \begin{bmatrix} 2 & 4 & 1 & 3 & 4 & 2 & 1 \\ 1 & 2 & 3 & 2 & 2 & 1 & 4 \\ 3 & 1 & 4 & 4 & 1 & 3 & 3 \\ 4 & 3 & 2 & 1 & 3 & 4 & 2 \end{bmatrix}$$

Let the **Mask** vector be given by

$$\text{Mask} = [0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0].$$

Now to obtain the new hypothesis from combining Hypotheses 1 and 2, we copy column 1, 2, 4, and 7 (the 0 cells) from Hypothesis 1 and the rest from Hypothesis 2 to form the New Hypothesis, shown next.

New Hypothesis after Combination Experiment:

$$\begin{array}{l} \text{Rule 1} \\ \text{Rule 2} \\ \text{Rule 3} \\ \text{Rule 4} \end{array} \begin{bmatrix} 1 & 3 & 1 & 2 & 4 & 2 & 3 \\ 3 & 2 & 3 & 1 & 2 & 1 & 1 \\ 4 & 4 & 4 & 3 & 1 & 3 & 2 \\ 2 & 1 & 2 & 4 & 3 & 4 & 4 \end{bmatrix}$$

In the experiment the new hypothesis replaces an existing low accuracy hypothesis with high probability. The selection of the hypothesis to be replaced is made using the SUS selection algorithm based on accuracy values. The selection process is formulated so that the least accurate hypothesis receives the greatest probability of selection.

Individual Hypothesis Experiments

An Individual Experiment is triggered whenever a Combination Experiment is not triggered. An Individual Experiment therefore occurs with a probability of $(1 - \text{Pr}(\text{Comb}))$. For each agent, the SUS selection algorithm is used to select one hypothesis to be modified based on the accuracy values of all the agent's hypotheses. The hypothesis with the lowest accuracy value will have a high probability of being selected. Each column of the 4 x 7 matrix of rules for the selected hypothesis is then modified. Each of the 7 columns in the hypothesis has a 50% chance of being

changed. So it is possible that more than one column may be changed at any given date t . Each column that is selected for changed is replaced with a new 4×1 column vector containing the numbers $[1 \ 2 \ 3 \ 4]$ where before substitution the vector is randomly shuffled. To implement this, we first construct a 4×24 matrix, \mathbf{A} , to store all the 24 possible permutations of the four numbers: $[1 \ 2 \ 3 \ 4]$. The columns of \mathbf{A} store each possible permutation of the vector $[1 \ 2 \ 3 \ 4]$. A random integer number is then drawn from $U[0,24]$ and this number identifies the location of a column in \mathbf{A} that will be used to replace the chosen column in the Individual Experiment. Recall that our identifiers for the four fuzzy states are the numbers (1,2,3,4).

Reference

Baker, J. E. "Reducing Bias and Inefficiency in the Selection Algorithm." In Proceedings of the Second International Conference on Genetic Algorithms and their Application, J.J. Grefenstette, ed. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates, 14-21 (1987).