Long correlations and Levy Models applied to the study of Memory effects in high frequency (tick) data

M.C. Mariani[1], I. Florescu[2], M.P. Beccar Varela[1], E. Ncheuguim[1]

[1] Department of Mathematical Sciences, New Mexico State
[2] Department of Mathematical Sciences, Stevens Institute of Technology

This work is devoted to the study of long correlations, memory effects and other statistical properties of high frequency (tick) data. We use a sample of 25 stocks for this purpose.

We verify that the behavior of the return is compatible with that of continuous time Levy processes. We also study the presence of memory effects and long-range correlations in the values of the return.

keywords: High frequency (tick) data, Stock Indices, Econophysics, Hurst analysis, Detrended Fluctuation Analysis

1. Introduction

In recent years there has been a growing literature in financial economics that analyzes the major stock indices in developed countries, see [1-5] and the references therein. One of the main problems is the analysis of the existence of long term or short term correlations in the behavior of financial markets. The statistical properties of the temporal series analyzing the evolution of the different markets have been of a great importance in the study of financial markets. The empirical characterization of stochastic processes usually requires the study of temporal correlations and the determination of asymptotic probability density distributions**.**

Studies that focus on a particular country index generally show that a long-term memory effect exists in those indices, see [6-7] and the references therein. The previous studies concentrated on daily data. We wish to verify if the same conclusion applies to high frequency data. Following this line, we analyze a sample of 26 stocks of trade-by-trade (tick) data for a very typical day (10-04-2007) devoid of any major events.

We find that all unit root tests performed rejected the existence of a unit root type nonstationarity. The p-values of the tests were all under 0.01.

We use Rescaled Range Analysis (R/S) and Detrended Fluctuation Analysis (DFA) methods to determine long-range correlations. Both methods characterize fractional behavior, but R/S analysis can yield more accurate results for small and stationary data sets and DFA analysis yields more accurate result for non stationary data sets. The exponents calculated are complementary and could serve as verification and comparison of the results; therefore, both methods are used.

We found evidence that even in an ordinary day without any notable information for about 75% of the market, the use of short term memory models is inappropriate. Specifically, in only 23% of the studied cases one of the tests performed did not reject the Gaussian hypothesis (no memory or very short term memory). There were no stocks for which both tests performed agreed that the data is Gaussian.

We conclude that stochastic volatility models, jump diffusion models and general Levy processes seem to be needed for the modeling of high frequency data in any situation.

2. R / S Analysis

Hurst developed the re-scaled range analysis (R/S analysis) [8, 9]. He observed many natural phenomena that followed a biased random walk, i.e., every phenomenon showed a pattern. He measured the trend using an exponent now labeled the Hurst exponent.

The procedure used to calculate R/S is as follows:

1. Let N be the length of time series (y_1, y_2, y_3…, y_N). The logarithmic ratio of the time series is obtained. The length of the new time series M(t) will be N – 1.

$$M(t) = \log\left(\frac{y_{t+1}}{y_t}\right), t = 1,2,....,(N-1)$$

2. The time series is then divided into m sub-series of length n. n represents the number of elements in the series and m represents the number of sub-series. Thus m * n = N -1. Each sub-series can be labeled as Q_a where a = 1, 2,…., m and each element in Q_a can be labeled as $L_{k,a}$ for k = 1,2,….,n.

3. For each Q_a, the average value is calculated:

$$Z_a = \frac{1}{n}\sum_{k=1}^{n} L_{k,a}$$

4. The cumulative deviation in each Q_a is calculated:

$$C_{k,a} = \sum_{j=1}^{k} \left(L_{j,a} - Z_a \right) \quad \text{k = 1,2,….,n}$$

5. Thus the range of each sub-series Q_a is given as:

$$R(Q_a) = \max(C_{k,a}) - \min(C_{k,a})$$

6. The standard deviation of each sub-series Q_a is calculated:

$$S(Q_a) = \sqrt{\left(\frac{1}{n}\right)\sum_{j=1}^{n} \left(L_{j,a} - Z_a \right)^2}$$

7. Each sub-series is normalized by dividing the range, $R(Q_a)$ by the standard deviation, $S(Q_a)$. The average value of R/S for sub-series of length n is obtained by:

$$\left(R/S \right)_n = \frac{1}{m} \sum_{a=1}^{m} \frac{R(Q_a)}{S(Q_a)}$$

8. Steps 2 through 7 are repeated for all possible values of n, thus obtaining the corresponding R/S values for each n.

The relationship between length of the sub-series, n and the rescaled range R/S is:

$$\frac{R}{S} = (c * n)^H$$

where R/S is the rescaled range, n is the length of the sub-series of the time series and H is the Hurst exponent. Taking logarithms yields:

$$\log\left(R/S \right) = H * (\log n + \log c)$$

9. An ordinary least squares regression is performed using log(R/S) as a dependent variable and log(n) as an independent variable. The slope of the equation is the estimate of the Hurst exponent, H.

If the value of H for the investigated time series is 0.5, then it implies that the time series follows a random walk, i.e. an independent process. For data series with long memory effects, H would lie between 0.5 and 1, or elements of the observation are dependent. This means that what happens now would have an impact on the future. This property of the time series is called persistent time series and this character enables prediction of any time series as it shows a trend. If H lies between 0 and 0.5, it implies that the time-series possess anti-persistent behavior (negative autocorrelation).

3. Detrended Fluctuation Analysis

The Detrended Fluctuation Analysis method ***(DFA)*** is an important technique in revealing long range correlations in non-stationary time series. This method was developed by Peng [10], and has been successfully applied to the study of cloud breaking, Latin-American market indices, DNA, cardiac dynamics, climatic studies, solid state physics, and economic series. The advantages of DFA over conventional methods are that it permits the detection of intrinsic self-similarity embedded in a seemingly non-stationary time series, and also avoids the spurious detection of apparent self-similarity, which may be an artifact of extrinsic trends.

First, the absolute value of M(t), i.e. the logarithmic returns of the indices calculated in the R/S analysis, is integrated:

$$y(t) = \sum_{i}^{t} |M(i)|$$

Then the integrated time series of length N is divided into m boxes of equal length n with no intersection between them. As the data is divided into equal-length intervals, there may be some left over at the end. In order to take account of these leftover values, the same procedure is repeated but beginning from the end, obtaining $2N$ / n boxes. Then, a least squares line is fitted to each box, representing the trend in each box, thus obtaining (y_n (t)). Finally the root mean square fluctuation is calculated using the formula:

$$F(n) = \sqrt{\frac{1}{2N}\sum_{t=1}^{2N}\left[y(t) - y_n(t)\right]^2}$$

This computation is repeated over all box sizes to characterize a relationship between the box size n and F (n). A linear relationship between the F(n) and n (i.e. box size) in a log-log plot reveals that the fluctuations can be characterized by a scaling exponent (α), the slope of the line relating log *F(n)* to log *n*.

For data series with no correlations or short-range correlation, alpha is expected to be 0.5. For data series with long-range power law correlations, alpha would lie between 0.5 and 1 and for power law anti-correlations; alpha would lie between 0 and 0.5. This method was used for the measure of correlations in highly traded financial series, and in the daily evolution of some of the most relevant indices.

4. Stationarity and Unit Root Test

To study the fractional behavior of a times series using the R/S or the DFA analysis, it is important to investigate whether the underlying time series is stationary or not. The first method is more appropriate when analyzing stationary data sets, whereas the second method is more appropriate for non-stationary data sets. In the economic literature we can find tests for a particular type of nonstationarity behavior: the unit-root nonstationarity. Assume that the process $\{y_t\}$ possesses a univariate autoregressive stochastic component of order p; that is $\{y_t\}$ obeys the equation

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \varepsilon_t ,$$

Where the ε_t corresponding to different values of t may be correlated but they are all stationary (i.e., have the same distribution for all *t*).

We define the roots of the AR (p) component as the solutions to the characteristic polynomial

$$\Phi(z) = z^p - a_1 z^{p-1} - a_2 z^{p-1} - ... - a_p .$$

The process $\{y_t\}$ is stationary if all the roots of the characteristic polynomial *Φ(z)* lie outside the unit circle. In this case shocks into the system will dissipate over time. If Φ(z) has at least one unit root, it is said to exhibit a unit-root type non-stationary behavior; the effect of shocks never dies out. We note that this is the only type of nonstationarity that can be formally tested.

5. Methods and data analysis

We study high frequency data for 26 stocks traded on NYSE during April 10 2007. We chose this particular day since we want to study the typical behavior of the equity data, during a day when there are no major events influencing the returns. We pick a sample of 26 highly traded stocks and for obvious reasons we call them Stock 1, Stock 2, …, Stock 26. Since we use every trade it is very common to find many consecutive trades at the same price. We cumulate all such consecutive trades into one data point since they do not indicate price movement. In this work, the stochastic variable analyzed is the continuously compounded return (r_t), defined as the difference of the logarithm of two consecutive equity prices:

$$r_t = \log(S_t) – \log(S_{t-1})$$

Due to the nature of the stock movement (only moves in $0.01 increments) the resulting values for the return are in fact discretized. There are many more data points where the stock changes just by one cent from transaction to transaction than points where the change in the stock price is higher. We can see this aspect of the data exemplified in Figure 1.

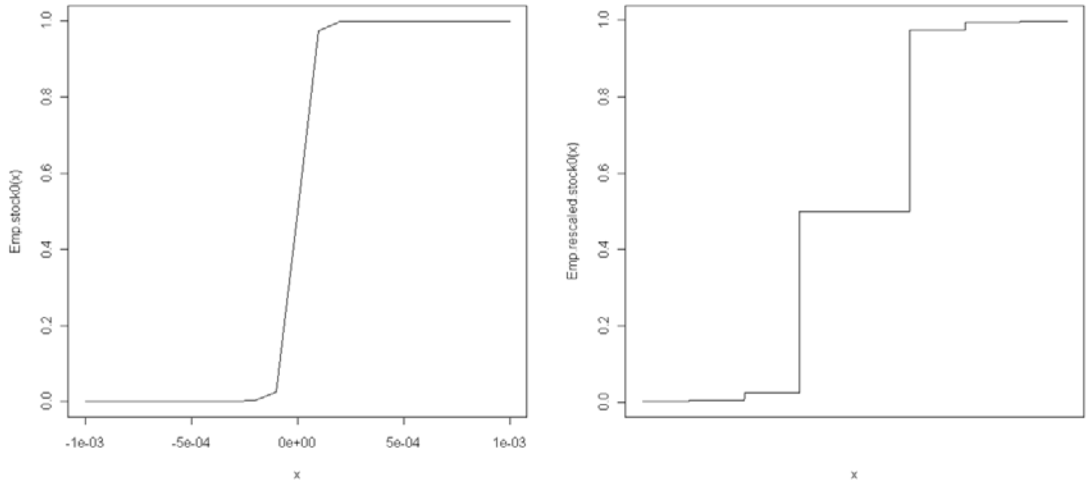

Figure 1: Plot of the empirical CDF of the returns for Stock 1. Left image contains the original CDF. The right image is the same empirical CDF but rescaled so that the discontinuities are clearly seen

The next images show the result obtained when comparing this empirical distribution function with the normal log-normal and logistic family of distributions. Additionally, we have compared with many other properly scaled families of distribution including Exponential, Gamma, and Weibull types. Of course all these are constructed assuming little or no memory in the dataset.

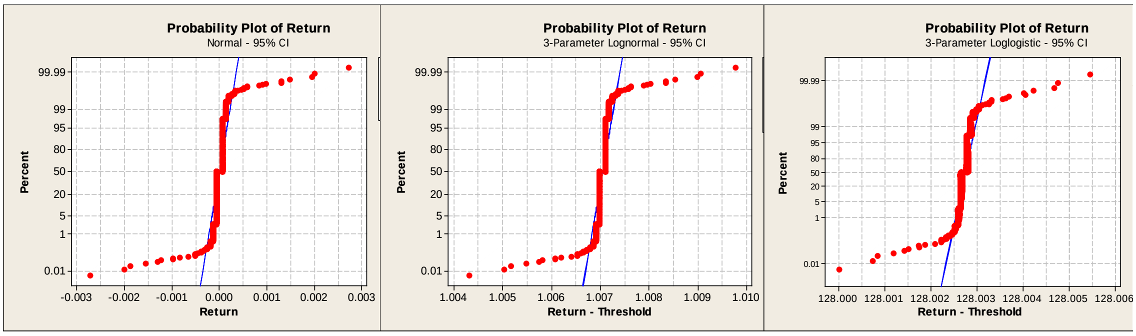

Figure 2: Quantile-Quantile plots of the empirical CDF of the returns for Stock 1 versus several candidate distributions. The plots and the numerical results reject all these traditional distributions.

The Table 1 below shows the results of the unit root test as well as the Hurst and DFA exponents:

Table 1

Data	ADF (Pvalue)	PP (p value)	KPSS (p value)	DFA	HURST
Stock 1	<0.01	<0.01	>0.1	0.525178 0.007037	0.561643 0.005423
Stock 2	<0.01	<0.01	>0.1	0.64812 0.01512	0.490789 0.006462
Stock 3	<0.01	<0.01	>0.1	0.66368 0.01465	0.628440 0.006138
Stock 4	<0.01	<0.01	>0.1	0.66969 0.01506	0.644534 0.005527
Stock 5	<0.01	<0.01	>0.1	0.65525 0.02916	0.65044 0.02908
Stock 6	<0.01	<0.01	>0.1	0.74206 0.01032	0.722893 0.008662
Stock 7	<0.01	<0.01	>0.1	0.50432 0.01212	0.644820 0.008521
Stock 8	<0.01	<0.01	>0.1	0.66184 0.01681	0.38046 0.01673
Stock 9	<0.01	<0.01	>0.1	0.72729 0.01383	0.635075 0.006374
Stock 10	<0.01	<0.01	0.07686	0.79322 0.01158	0.654970 0.006413
Stock 11	<0.01	<0.01	>0.1	0.322432 0.007075	0.52485 0.01265
Stock 12	<0.01	<0.01	>0.1	0.70352 0.01429	0.596178 0.007172
Stock 13	<0.01	<0.01	>0.1	0.74889 0.02081	0.58279 0.00825
Stock 14	<0.01	<0.01	>0.1	0.70976 0.01062	0.578053 0.007177
Stock 15	<0.01	<0.01	>0.1	0.76746 0.01029	0.588555 0.004527
Stock 16	<0.01	<0.01	>0.1	0.62549	0.61023

				0.01554	0.01083
Stock 17	<0.01	<0.01	>0.1	0.80534 0.02432	0.591336 0.006912
Stock 18	<0.01	<0.01	0.076	0.69134 0.01336	0.596003 0.001927
Stock 19	<0.01	<0.01	>0.1	0.678050 0.009018	0.596190 0.005278
Stock 20	<0.01	<0.01	>0.1	0.48603 0.01462	0.59426 0.01829
Stock 21	<0.01	<0.01	>0.1	0.65553 0.02517	0.50115 0.01086
Stock 22	<0.01	<0.01	>0.1	0.70807 0.01081	0.552367 0.009506
Stock 23	<0.01	<0.01	>0.1	0.717223 0.009553	0.594051 0.006709
Stock 24	<0.01	<0.01	>0.1	0.45403 0.01370	0.37129 0.02267
Stock 25	<0.01	<0.01	0.02718	0.63043 0.01200	0.646725 0.005784
Stock 26	<0.01	<0.01	>0.1	0.59568 0.01464	0.51591 0.01586

ADF= Augmented Dickey-Fuller Test for unit root stationarity
PP= Phillips-Perron Unit Root Test
KPSS= Kwiatkowski-Phillips-Schmidt-Shin Test for unit root Stationarity
DFA= Detrended Fluctuation Analysis
Hurst= Rescale Range Analysis
For the ADF and the PP the null hypothesis is the non-stationarity, and for the KPSS the null hypothesis is stationarity .
With two small exceptions the tests reject the unit-root type nonstationarity.

It is worth mentioning that while the stationarity tests reject the presence of the unit-root in the characteristic polynomial that does not necessarily mean that the data is stationary, only that the particular type of nonstationarity indicated by the unit root is absent. For this reason we proceed with both tests even though conventional wisdom would recommend the use of the Hurst analysis at this point.

The Figure 3 below show the plot of

$$\log\left(R/S\right) = H * (\log n + \log c) \quad \text{for Hurst}$$

and the plot of

$$\log F(n) = \alpha^* \log n + log\ c \quad \text{for DFA.}$$

for four stocks. The plots for the entire sample of 26 stocks can be obtained at:

www.math.stevens.edu/~ifloresc/fractional.html

Points close to a straight line indicate good parameter estimators.

6. Results and Discussion

The estimated values for the slopes are presented in the last two columns of Table 1. With one exception the results obtained using the two methods agree. 19 out of the 26 equity data analyzed (or about 73% of the data) exhibited long memory effects which were recognized by both DFA and R/S methods. For 6 out of the 26 (about 23%) one of the two tests did not indicate correlations in the data. In one case (Stock 7) the results were contradictory; both tests indicated the presence of the long memory effects, however, while R/S indicated a persistent behavior, the DFA shows an anti-persistent activity (negative correlation). Of the 19 stocks that show definite evidence of long memory effects 18 show a persistent and only 1 an anti-persistent activity.

We found evidence that even in an ordinary day without any notable information for about 75% of the market, the use of short term memory models is inappropriate. We conclude that stochastic volatility models, jump diffusion models and general Levy processes seem to be needed for the modeling of high frequency data in any situation.

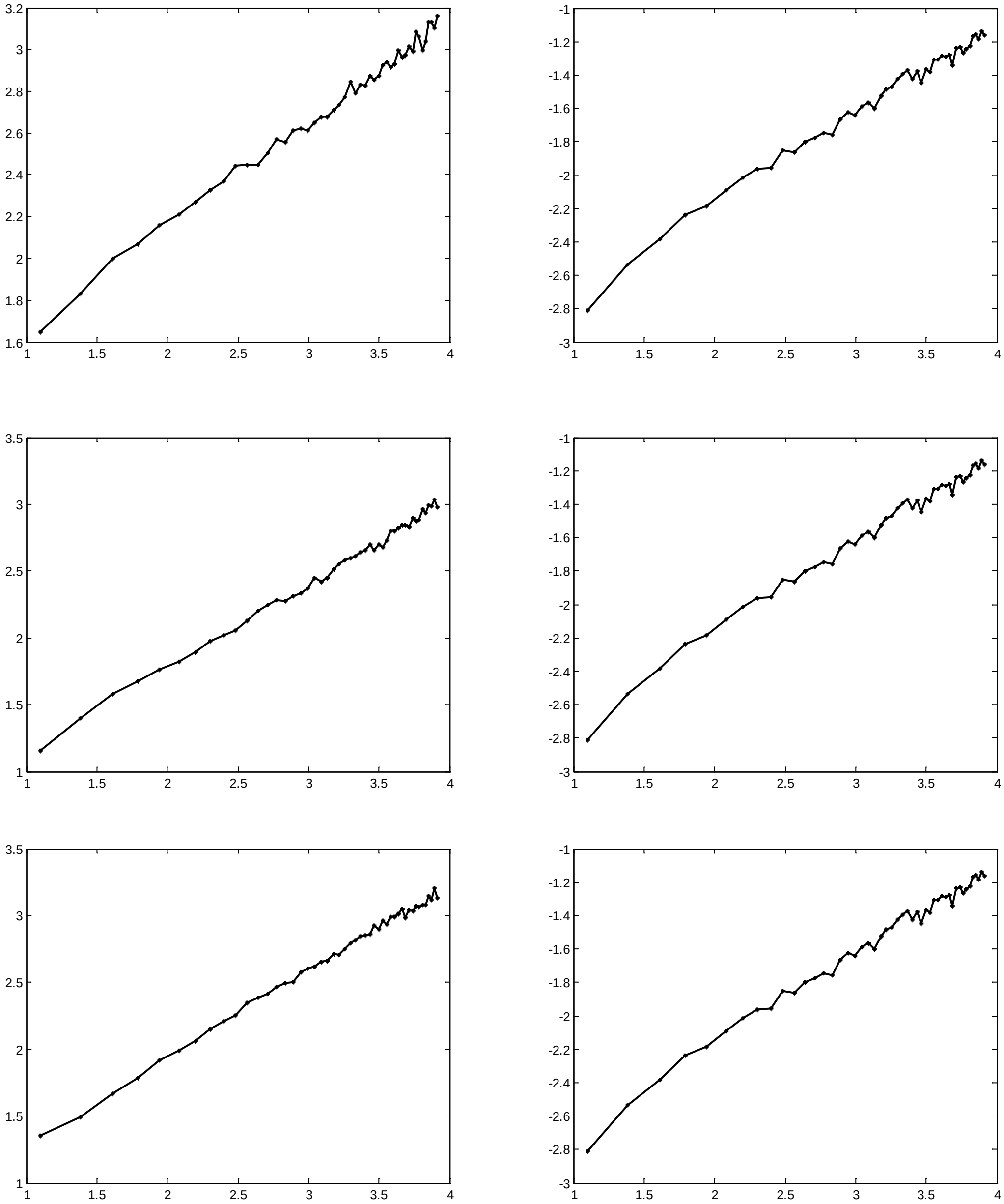

Figure 3: Hurst and DFA regression plots for a sample of three stocks. Plots on the left depict the Hurst method while the plots on the right show the results obtained using the DFA method.

Addresses for correspondence:

Dr. Maria Christina Mariani
Department of Mathematical Sciences, New Mexico State University
Sciences Hall 236
Las Cruces, NM 88003-8001 USA
Email: mmariani@nmsu.edu

Dr. Ionut Florescu
Department of Mathematical Sciences, Stevens Institute of Technology
Castle Point on Hudson
Hoboken, NJ 07030   USA
Email: Ionut.Florescu@stevens.edu

Acknowledgements

This work has been partially supported by Minigrant, College of Arts & Sciences NMSU.

References

[1] R. N. Mantegna, H. E. Stanley, An Introduction to Econophysics: Correlations and Complexity in Finance, Cambridge University Press, Cambridge 1999.

[2] H. E. Stanley, L.A.N. Amaral, D. Canning, P. Gopikrishnan, Y. Lee, Y. Liu, "Econophysics: Can Physicists Contribute to the Science of Economics?" [Proc. 1998 Econophysics Workshop], Physica A 269, (1999) 156-169.

[3] M. Ausloos, N. Vandewalle, Ph. Boveroux, A. Minguet, K. Ivanova, "Applications of statistical physics to economic and financial topics"**,** Physica A 274, (1999) 229-240.

[4] J.-Ph. Bouchaud, M. Potters, Théorie des riques financiers, Alea-Saclay/Eyrolles, Paris, 1997.

[5] R. N. Mantegna, H. E. Stanley, "Scaling Behaviour in the Dynamics of an Economic Index," Nature 376, (1995) 46-49.

[6] Y.H. Liu, P. Cizeau, M. Meyer, C.K. Peng, H.E. Stanley, "Quantification of Correlations in Economic Time Series," Physica A 245 (1997) 437 - 440.

[7] M. C. Mariani, J.D. Libbin et all, "Long correlations and Normalized Truncated Levy Models applied to the study of Indian Market Indices in comparison with other emerging markets"" to appear in Physica A.

[8] H. E. Hurst, Long term storage of reservoirs, Trans. Am. Soc. Civil Eng. 116, (1950) 770 – 808.

[9] B. B. Mandelbrot, Fractals and Scaling in finance (Springer, New York, 1997 edition).

[10] C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, "Mosaic organization of DNA nucleotides", Phys. Rev. E 49 (1994) 1685-1689.