THE UNIVERSITY OF RHODE ISLAND

University of Rhode Island
DigitalCommons@URI

Environmental and Natural Resource Economics Faculty Publications

Environmental and Natural Resource Economics

2016

Time Scale and Fractionality in Financial Time Series

Thomas W. Sproul

Follow this and additional works at: https://digitalcommons.uri.edu/enre_facpubs

The University of Rhode Island Faculty have made this article openly available.
Please let us know how Open Access to this research benefits you.

This is a pre-publication author manuscript of the final, published article.

Terms of Use
This article is made available under the terms and conditions applicable towards Open Access Policy Articles, as set forth in our Terms of Use.

Time Scale and Fractionality in Financial Time Series

Structured Abstract

Purpose: Turvey (2007, Physica A) introduced a scaled variance ratio procedure for testing the random walk hypothesis (RWH) for financial time series by estimating Hurst coefficients for a fractional Brownian motion model of asset prices. This article extends his work by making the estimation procedure robust to heteroskedasticity and by addressing the multiple hypothesis testing problem.

Methodology: Unbiased, heteroskedasticity consistent, variance ratio estimates are calculated for end-of-day price data for eight time lags over 12 agricultural commodity futures (front month) and 40 U.S. equities from 2000-2014. A bootstrapped stepdown procedure is used to obtain appropriate statistical confidence for the multiplicity of hypothesis tests. The variance ratio approach is compared against regression based testing for fractionality.

Findings: Failing to account for bias, heteroskedasticity, and multiplicity of testing can lead to large numbers of erroneous rejections of the null hypothesis of efficient markets following an independent random walk. Even with these adjustments, a few futures contracts significantly violate independence for short lags at the 99% level, and a number of equities/lags violate independence at the 95% level. When testing at the asset-level, futures prices are found not to contain fractional properties, while some equities do.

Research limitations: Only a subsample of futures and equities, and only a limited number of lags, are evaluated. It is possible that multiplicity adjustments for larger numbers of tests would result in fewer rejections of independence.

Value: This paper provides empirical evidence that violations of the random walk hypothesis for financial time series are likely to exist, but are perhaps less common than previously thought.

Keywords: financial time series, fractional Brownian motion, heteroskedasticity, multiple hypothesis testing, random walk, variance ratio

JEL Codes: Q14, G13, G12

1 Introduction

In his seminal paper on the efficient markets hypothesis (EMH), Fama (1970) stated that markets fully reflecting all available information is so general a definition of efficiency as to be empirically untestable. He therefore defined efficiency in terms of price formation according to a "fair game" model, in which returns based on current information follow a submartingale. A direct implication of this definition is that prices must follow a random walk if both short and long positions are allowed on all assets, the so-called random walk hypothesis (RWH). However, broad dissent arose in the 1980s with Grossman and Stiglitz (1980) identifying equilibrium disequilibrium in the form of rents to information for arbitrageurs, both Shiller (1981) and DeBondt and Thaler (1985) identifying over-reaction, and Lo and Mackinlay (1988) introducing variance ratio testing to debunk RWH econometrically.

Since that time, further evidence against EMH has accumulated along the lines of supply-and-demand imbalances (Hirshleifer, 1989, 1990; Bessembinder, 1992) and limits to arbitrage (Shleifer and Vishny, 1997). Despite these challenges, Fama (1998) concluded that market efficiency "survives the challenge from the literature" because the anomalies identified are essentially chance results; namely, under-reaction/overreaction and momentum/mean-reversion are about equally common. More recently, the mid-2000s spike in commodity prices concurrent with the development of commodity index investing led to claims of a price bubble (Masters, 2008) and a large number of articles evaluating efficiency of futures markets, with mixed evidence. A series of papers by Scott Irwin and coauthors (e.g., Irwin, 2013) dismissed these claims based on limited evidence across a wide range of approaches, and there continues to be conflicting evidence. Acharya et al. (2013) developed a model in which capital constrained speculators lead to *increased* hedging costs and verified it using 30 years of oil and gas futures, whereas Brunetti and Reiffen (2014) showed that commodity index traders actually *lower* hedging costs using a proprietary daily data set from the Commodity Futures Trading Commission (CFTC).

In contrast to futures markets, findings of inefficiencies in equities and options markets are not as mixed in recent years, especially along the lines of limits to arbitrage and financial institutions/infrastructure as key drivers. Bollen and Whaley (2004) found evidence that implied volatility is a function of net buying pressure from public order flow in index options, and Constantinides et al. (2009) identified excess returns in index put options that violate first order dominance. On the supply side, Duffie (2010) showed that slow-moving capital leads to commonplace, temporary price imbalances (e.g., up to a month) in equities but not longer term imbalances because market participants have time to overcome search costs and short-term capital constraints. On the demand side, Hong et al. (2012) showed that arbitrageurs amplify economic shocks in equities through the unwinding of speculative short positions following good news, and Cao et al. (2013) found evidence that some hedge funds are able to forecast market liquidity.

To summarize, the literature suggests substantial evidence against the stronger forms of EMH, but violations may be more consistently found in equities than in future markets. This is intuitively reasonable based on liquidity alone, but does require explicit testing. The time scale of these violations is also in question, as short- and medium-term violations (e.g., as in Asness et al., 2013) appear to be most common. Turvey (2007, herein Turvey) addressed the "memory" inherent in agricultural commodity prices by

extending the fractional Brownian motion model of Mandelbrot and van Ness (1968) into a scaled variance ratio estimator of the Hurst coefficient and demonstrating superior Monte Carlo performance against their original rescaled range ("R/S") statistic. Turvey used simulated confidence intervals for the null hypothesis of $\hat{H} = 0.5$ to show that agricultural commodity prices for the period 1996-2001 do not, in general, allow for rejection of the null hypothesis of a geometric Brownian motion, but that some individual commodities do exhibit fractional properties.

A key caveat is in order. In light of the large quantity of financial data available, and the multiplicity of testing conducted by financial researchers, there are legitimate concerns of data-snooping bias and related problems when the literature is evaluated as a whole. Data-snooping bias is generally defined as statistical inference performed after the researcher has inspected the data (it may include Leamer's (1978) "specification searches," for example). However there is also a similar meta problem that can arise when the researcher begins to inspect data after many other researchers have already performed mass inference testing on the same data. Fama (1998), Turvey, and Irwin and Sanders (2011) all implied these concerns in the manner of their evaluation of literature and results, but do not state them explicitly, whereas Harvey et al. (2014), in a recent NBER working paper, advanced the notion that standards for new findings in finance should be revised to only accept significance findings for t-ratios greater than 3.0. Naturally, this solution offers some protection against both classes of problem, but it may not go far enough against within-researcher data-snooping bias because of the sheer sample sizes available in many financial time series. Of course, within-researcher data-snooping concerns are not new, as improvements on the basic Bonferroni correction for multiple hypothesis testing in the last two decades have become seminal papers in their own right. These papers include Benjamini and Hochberg (1995) developing a procedure for controlling the false discovery rate, Benjamini and Yekutieli (2001) refining this procedure for dependence between tests, and Romano and Wolf (2005a, 2005b) proving and demonstrating a subsampling method for stepwise control of the familywise error rate, with both improved finite sample power and desirable asymptotic properties.

In this article, we extend Turvey's approach to be robust to both heterskedasticity and non-normal innovations by relying on the asymptotic results of Lo and Mackinlay (1988)(herein, LM). To address concerns regarding the large number of hypothesis tests inherent in this type of analysis, we extend the procedure of Romano and Wolf (2005b) to the variance ratio testing conducted herein using a block bootstrap. Results of the stepdown procedure are then held to the standard of Harvey et al. (2014) to protect against meta-level data-snooping across studies in finance. By combining these approaches, we are able to compare and contrast variance ratio testing by asset and time lag against Turvey's regression-based approach to identify fractional properties at the asset level.

We find that a properly adjusted variance ratio testing procedure fails to reject random walks for nearly all agricultural commodity futures prices (and lags) for the period 2000-2014 (and fails to reject all of them using a 3σ standard), and our extended version of Turvey's approach does not reject random walks for any of the commodity futures price series tested. Variance ratio testing of the same time series for U.S. equities leads to no rejections at suitable levels because many equities prices/lags exhibit adjusted *p*-values below 0.05, but none below 0.01. However, when testing our extension of

Turvey's procedure against equities at the asset level, a substantial percentage of equities tested show strong fractional properties, significant at the 99.9% level. Since it is well-known that variance ratio testing can be overly sensitive to short-term dependence (Lo, 1991), we take this pattern of results as evidence of long-term memory in the time series of some equities prices.

Overall, it is shown that failing to adjust for heteroskedasticity and for the multiplicity of hypothesis tests can lead to too many rejections of the independence null, and to substantial statistical overconfidence in the strength of the rejections (because unadjusted *p*-values will indicate a likelihood of results due only to random chance which is far too small). In doing so, we provide rigorous evidence in support of the findings of Irwin and Sanders (2011) that commodity futures markets are likely to be weak-form efficient. Our findings are similar for the Dow Jones (DIA) and S&P 500 (SPY) exchange traded funds (ETFs), and for the majority of Dow stocks. However, we do find significant evidence of long-term memory in a subset of Dow stocks, consistent with the original results of LM.

The remainder of this article is organized as follows. The next section describes the theory and econometrics behind the results. It gives a brief review of the theory behind fractional Brownian motion and the role of the Hurst parameter in determining memory, or lack thereof, in a time series. This is followed by a review of the method of Turvey and its extension to a heteroskedasticity-consistent form using the asymptotic results of LM, and a review of the Romano and Wolf stepwise subsampling procedure for controlling FWER. Section 3 describes our data sources and Section 4 presents the results and demonstrates the impact of correcting for the multiple-hypothesis testing problem. Section 5 discusses implications for future research, and concludes.

2 On Variance Ratio Tests, "Scaled" and Multiple

2.1 Fractional Brownian Motion and the Scaled Variance Ratio

In the tradition of the Black-Scholes-Merton model, asset prices can be modeled as following a geometric Brownian motion (gBm), so that the log prices follow a standard (arithmetic) Brownian motion (Bm). Turvey utilizes the fractional Brownian motion (fBm) as a generalization of Bm that (i) incorporates Bm as a special case, and (ii) provides a conveniently consolidated alternative hypothesis for empirical testing. For our purposes, the key feature of fBm is its autocorrelation structure. Namely, if x is an fBm process, it is a continuous time Gaussian process with mean zero and autocovariance

$$\mathrm{E}\left[x(t+\Delta t)x(t)\right]=\frac{1}{2}\sigma^2\left((t+\Delta t)^{2H}-t^{2H}-\Delta t^{2H}\right) \tag{1}$$

which reduces to zero (and the standard Bm) for $H=0.5$. For our purposes, the key feature of fBm is the growth of variance over time, given by

$$\mathrm{E}\left[(x(t+\Delta t)-x(t))^2\right]=\sigma^2(\Delta t)^{2H}, \tag{2}$$

which collapses to $\sigma^2 t$ when $H=0.5$ (the standard Bm exhibits variance linear in time). Critically, $H<0.5$ will cause the series to exhibit mean-reversion and variance growing

more slowly than time, while $H > 0.5$ will cause the series to exhibit momentum or trend-following properties with variance growing more rapidly than time. In both cases, the fractional process will exhibit long-term memory, exhibited by slower decay of the autocorrelation coefficients for large lags than is found in short-term dependent processes (like an AR(1) process, as noted in Lo (1991)).

We thank a referee for pointing out that fBm generalizes the standard Bm in some sense, but it is not the only generalization possible. Furthermore, fBm with $H \neq 0.5$ is not a semi-martingale, which means that typically the no-arbitrage property is ruled out – a feature which may not be desirable in asset pricing models. However, Bender et al. (2007) show that no arbitrage pricing is possible for a broad class of portfolios when a mixed-fBm model is considered, one in which innovations follow a mixture distribution with both an fBm and standard Brownian component. Despite these features of the more generalized modeling problem, our presentation here is for consistency with that developed in Turvey.

An immediate consequence of Eq. 2 is that the ratio of the *k*-period variance to the one-period variance is then given by

$$\frac{\mathrm{E}\left[(x(t+k)-x(t))^2\right]}{\sigma^2} = k^{2H}. \tag{3}$$

Letting $\widehat{VR}(k)$ denote the variance ratio on the left-hand side of Eq. 3, estimated from data, yields an estimate of the Hurst parameter: $\hat{H} = \ln \widehat{VR} / 2 \ln k$. Clearly, $\hat{H} = 0.5$ in expectation for a standard Bm, but it is necessarily subject to sampling variability.

Turvey recognizes that more stable estimates of $\hat{H}$ over the full sample price series can be generated by regression analysis using estimated variance ratios for a number of lags. In addition to stability, this procedure also has the benefit of isolating systematic deviations from gBm across all lags, while being less sensitive to autocorrelation coefficients that are only distorted for shorter lags, as might arise from a short-term dependent process. The estimating equation is:

$$\ln \widehat{VR}(k) = \beta_0 + \beta_1 \cdot \ln k + \epsilon \tag{4}$$

where the null hypothesis is $H_0 : \beta_0 = 0, \beta_1 = 1$ and where $\hat{H} = 0.5\hat{\beta}_1$. Turvey goes on to show that confidence intervals for rejection of standard Bm as the null hypothesis $\left(H_0 : \hat{H} = 0.5\right)$ can be simulated as a function of the lag, *k*, and sample size, and that this procedure outperforms other established procedures for estimating Hurst, especially rescaled range analysis.

We also thank a referee for pointing out the advantage of the original Hurst exponent, estimated from the R/S rescaled range statistic. Namely, it is well-defined for all types of distributions, including those with fat-tailed innovations and infinite variances. We justify our approach in terms of the discussion provided by Lo (1991), who points out that while the R/S statistic may be robust to a broader class of processes, the R/S statistic is sensitive to short-term dependence, which ultimately must be corrected

using knowledge of autocorrelations. Lo (1991) also points out that the strong-mixing assumptions supporting the LM heteroskedasticity consistent variance ratio estimator (which we use, see Section 2.2 below) do not support infinite variance processes but they do allow for unconditional leptokurtosis via time-varying conditional heteroskedasticity.

The key innovations in this paper involve extending Turvey's method by combining a block bootstrap, weighted least squares, and a stepdown procedure to control the familywise error rate (FWER) explicitly in the context of multiple hypothesis testing. The methods are shown to be complementary to the variance ratio testing approach in understanding the nature of various financial time series. We do not show whether extension of the original R/S statistic to multiple testing is similarly straightforward, but it is likely that some adaptation of the R/S statistic to our framework is possible.

2.2 Asymptotic Heteroskedasticity-Consistent (Scaled) Variance Ratio Tests

A key challenge in extending Turvey's method more broadly is accounting for heteroskedasticity, given ample evidence in the literature of persistence, clustering and autocorrelation of volatility for financial time series. LM work with a standardized variance ratio, $VR_1(k) = VR(k)/k$, and derive a test statistic that is asymptotically normal, heteroskedasticity-consistent, and robust to non-normality of innovations, while also having good finite sample performance through the use of overlapping periods for estimating sample autocorrelations.

The variance ratio test statistic of LM considers the compound null hypothesis of a random walk with: (i) uncorrelated increments, (ii) sample autocorrelations asymptotically uncorrelated with one another, (iii) finite variance, and (iv) a mixing condition limiting the maximum amount of dependence and heterogeneity while still inducing the Law of Large Numbers and the Central Limit Theorem (see White, 1984). Construction of the test statistic for $\widehat{VR_1}(k)$ proceeds as follows, with notation adapted to be internally consistent herein. First, let $nk = T$ and define the variance ratio estimator to be the ratio of unbiased variance estimates for lags *k* and *1*, estimated from overlapping periods in the case of *k*:

$$\widehat{VR_1}(k) = \frac{(nk-1)\sum_{t=k}^{nk}\left(p_t - p_{t-k} - k\hat{\mu}\right)^2}{k(nk-k+1)\left(1-\frac{k}{nk}\right)\sum_{t=1}^{nk}\left(p_t - p_{t-1} - \hat{\mu}\right)^2}, \tag{5}$$

where the coefficient on the denominator represents bias correction of the sample *k*-lag variance due to overlapping periods. This variance ratio estimator is still necessarily biased due to Jensen's inequality, but Lo and Mackinlay (1989) show that its finite-sample properties are close to their asymptotic limits. To correct for heteroskedasticity, let $\hat{\rho}_j$ be the sample autocorrelation coefficient at lag *j*, and note that the variance ratio has the asymptotic relationship:

$$\widehat{VR_1}(k) \overset{a}{=} 1 + 2\sum_{j=1}^{k-1}\left(1-\frac{j}{k}\right)\hat{\rho}_j\,. \tag{6}$$

Letting $nk = T$, a heteroskedasticity-consistent estimator for the variance of $\hat{\rho}_j$ is:

$$\hat{\delta}_j = \frac{nk \sum_{t=j+1}^{nk} \left(p_t - p_{t-1} - \hat{\mu}\right)^2 \left(p_{t-j} - p_{t-j-1} - \hat{\mu}\right)^2}{\left(\sum_{t=1}^{nk} \left(p_t - p_{t-1} - \hat{\mu}\right)^2\right)^2} \tag{7}$$

where $\hat{\mu} = (p_T - p_0)/T$ is the mean drift of the sample. Accordingly, the asymptotic variance of $\widehat{VR}_1(k)$ is estimated in a heterskedasticity-consistent fashion by:

$$\hat{\theta}(k) = 4 \sum_{j=1}^{k-1} \left(1 - \frac{j}{k}\right)^2 \hat{\delta}_j . \tag{8}$$

The test statistic of LM is then simply:

$$\varphi^*(k) = \frac{\sqrt{nk}\left(\widehat{VR}_1(k) - 1\right)}{\sqrt{\hat{\theta}(k)}} \overset{a}{\sim} \mathcal{N}(0,1). \tag{9}$$

This test-statistic can then be used to generate *p*-values for standard hypothesis testing, though the applied researcher needs to take care when testing a large number of hypotheses, such as when a panel of asset prices are tested across multiple time lags. In the stepdown method described below, either the test-statistics or the associated *p*-values can be used to obtain equivalent results, subject to the caveat that the stepdown procedure is one-sided, so test-statistics must be converted to absolute values for a two-sided test.

2.3 Multiple Hypothesis Variance Ratio Testing

As will be shown below, we intend to test variance ratios for financial time series covering (12 futures contracts + 40 equities) × 8 lags, resulting in $J = 416$ hypothesis tests. For this many tests, failing to correct for the number of tests will lead to an inflated Type I error rate (too many false rejections or "false discoveries"). For example, consider a single test with $\alpha = 0.001$, which has a 0.1% chance of incorrectly rejecting a true null hypothesis as false. If 416 such tests are conducted, then the probability of at least one incorrect rejection is given by $1 - (1 - \alpha)^{416} = 34.05\%$. Correcting this problem explicitly means controlling the familywise error rate (FWER), which is the probability of at least one false discovery among all the hypotheses when performing multiple hypothesis tests. The most basic (worst-case) method to control FWER is the Bonferroni method of rejecting the null only for *p*-values less than $1/J$. Unfortunately, this method is overly conservative in many cases, leading to too many failures to reject.

To obtain additional power over Bonferroni, we use the method of Romano and Wolf (2005b), herein RW, who demonstrate a bootstrapped stepdown procedure based on the empirical distribution of the maximal test statistic under the null hypothesis. This

procedure not only controls for non-normality of innovations and heteroskedasticity, two well-documented features of financial time series, it also increases the power of multiple testing by taking advantage of the dependence structure between tests. In particular, a block bootstrap (Lahiri, 2003) to preserve within-period cross-correlations between equities adds additional power since the equities examined herein are highly correlated due to their membership in the Dow average. To be clear, cross-correlations of asset returns within time periods, e.g., due to macroeconomic or other factors, in no way violates RWH, which is only concerned with autocorrelation within single time series. Thus, the block-bootstrap helps generate the sampling distribution under the null hypothesis of no serial correlation within assets.

The algorithm proceeds as follows. Let $\tau_j = 1 - \tilde{p}_j$ (1 minus the estimated, two-sided *p*-value) be the test statistic for hypothesis H_j, and let $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_J$ be the sorted test statistics over all hypotheses. Test the intersection hypothesis, $H_J = \bigcap H_j$. If H_J is rejected, set aside H_1 as rejected; otherwise, accept all hypotheses. Following rejection of an individual hypothesis, repeat this step, forming and testing a new intersection hypothesis from the remaining individual hypotheses. The procedure ends as soon as one intersection hypothesis is accepted. Testing an intersection hypothesis is performed by estimating the distribution of the maximum of the test statistics $\tau_1,\ldots,\tau_J$. The intersection hypothesis is rejected if $\tau_1 \geq x^*$, where the critical value is estimated according to:

$$x^* = \inf\left\{x : \Pr\left(\max\left\{\tau_1,\ldots,\tau_J\right\} \leq x\right) \geq 1-\alpha\right\} \tag{10}$$

using the bootstrapped data. Here, α is the desired FWER. Following RW, we use $B = 10{,}000$ bootstrap replications for each test, where the bootstrapped sample has the same number of observations as the original sample, and where the block bootstrap indicates permutations of "rows" consisting of cross-sectional asset returns at time *t*. In this manner, the block bootstrap is used as a form of permutation testing, that (i) remains robust to realized heteroskedasticity (and non-normality), while (ii) allowing for construction of the distribution of test statistics under the null of no auto-correlation within assets.

2.4 A Combined Approach

There is an obvious connection between the one-off estimate of $\hat{H}$ and the estimated variance ratio: this estimate of the Hurst parameter is a monotone increasing transform of the variance ratio, and it is trivial to show that tests of the estimated Hurst parameter are equivalent to any other variance ratio test using a monotone increasing transform, so long as the underlying variance ratio estimates agree. Figure 1 below shows the 95%, 99%, and 99.9% asymptotic confidence intervals for the LM asymptotic variance ratio, alongside the same intervals for $\hat{H}$ (the heteroskedasticity-consistent version is not shown; it is similar, but the exact curves are necessarily data dependent). For single hypothesis tests, the beauty of this comparison is that the monotone increasing transform preserves the confidence intervals, so the applied researcher can take advantage of the

asymptotic results of LM for hypothesis testing in large samples without relying on simulated confidence intervals.

[Figure 1 about here]

On the other hand, multiple hypothesis testing (a common situation when dealing with financial time series) is not nearly so straightforward, as formalized procedures must be used to control FWER. In the case of variance ratio testing in particular, even when multiple hypothesis testing is executed correctly, the pattern of null rejections may be inconsistent across assets and lags leaving the researcher to guess at the underlying causes. Adding to this problem is the fact that variance ratio tests provide increasingly weak identification as the number of lags grows. Consider the simple example of estimating $\widehat{VR}_{50}(k)$ from Eq. 6: $\hat{\rho}_1$ gets 49 times the weight of $\hat{\rho}_{49}$ in this estimate, meaning that even large lags generate variance ratios that are primarily concerned with low-order autocorrelations. This feature of variance ratio testing means, among other things, that estimated variance ratios may be perfectly consistent with an fBm process, but still fail to reject the null at larger time lags due to lack of power. This lack of power can arise both from construction of the test statistic, which places lower weights on long-lag autocorrelations relative to shorter lags, and from the subtle structure of the sample autocovariance matrix generated by an fBm process.

The essence of Turvey's method is recognizing that testing a more restrictive alternative hypothesis can lead to more useful insights as to the nature of financial time series behavior. Through regression, it can be estimated whether variance ratios grow in a fractional manner as described by fBm, and whether they have a fixed component indicative of heteroskedasticity or other non-independent dynamics not fully captured by the fBm model. These estimates are much more useful *a posteriori* than the individual variance ratio tests because they characterize the entire time series, and because individual variance ratio tests may be overly sensitive to short-term dependence.

We propose the following extension of Turvey's method:

1. Estimate $\widehat{VR}(k)$ for each asset × lag combination, using the bias-corrected method of LM, but without dividing by *k*.
2. Estimate $\widehat{VR}(k)$ again for each of the block bootstrapped permutations, to obtain the sample standard deviation of $\ln\widehat{VR}(k)$, conditional on lag.
3. Estimate $\hat{\beta}_0, \hat{\beta}_1$ via Eq. 4 on the original data using weighted least squares (WLS), with sample size (if it varies) and sample standard deviation to generate the weights (see below).
4. Repeat step 3 for the bootstrapped distribution of $\widehat{VR}(k)$.
5. Choose the desired FWER (α) and apply the RW stepdown procedure using $\tau_j = 1 - \tilde{p}_j$ for the test statistics, where $\tilde{p}_j$ is the two-sided *p*-value estimated via WLS.

The above algorithm generates estimates of the Turvey regression equation (and consequently, of $\hat{H} = 0.5 \cdot \hat{\beta}_1$) that are robust to heteroskedasticity innovations, as well as correctly and explicitly tested within a multiple hypothesis testing framework. The weighted least squares estimates are more efficient than standard OLS because the structure of heteroskedasticity under the null hypothesis can be directly controlled.

As mentioned in the Data section below, there can be cases of missing data for some of the assets being tested. Letting T_{ik} denote the number of observed returns for asset i at lag k, and T_{ikm} denoting the same for the mth bootstrapped sample drawn, the conditional sample standard deviation for WLS is given by:

$$\frac{1}{\sqrt{w_{ik}}} = \frac{\sqrt{T_{ik}}}{\hat{\sigma}\left(\sqrt{T_{ikm}} \cdot \ln \widehat{VR}_{ikm}(k)\right)} \quad . \tag{11}$$

The next section discusses the data used in this study. The following section presents the results, comparing the variance ratio testing approach against the extended (Turvey) regression approach in the context of explicit multiple hypothesis testing.

3 Data

All data are sourced from Quandl (quandl.com) via their API. End of day agricultural commodity futures closing prices are from the Chicago Board of Trade (CME/CBOT) and the InterContinental Exchange (ICE). We use rolling front-month expiries only, following Turvey. The CME contracts evaluated are corn (C), oats (O), soybeans (S), and wheat (W), as well as feeder cattle (FC), live cattle (LC), and lean hogs (LN). The ICE contracts are cocoa (CC), coffee (KC), cotton (CT), orange juice (OJ), and sugar (SB). Following Turvey, the rolling futures prices are not adjusted to address splicing bias, which might arise when lagged returns are calculated across contract expiries. We thank a referee for pointing this out, but note that the combination (i) of a wide range of lags in our analysis, (ii) relatively close spacing in time of contract expiries, and (iii) liquidity issues in non-front-month contracts, creates substantial challenges in adapting the computation of variance ratios to adjust for splicing bias explicitly. In the interest of brevity, we leave this extension as an area for future research.

End of day equities closing prices come from Yahoo Finance and are adjusted for splits and dividends. Data are obtained for the DIA (Dow) and SPY (S&P 500) ETFs, as well as the 38 stocks making up the Dow Jones Industrial Average (DJIA) at any time in the sample period. The stocks are the current Dow 30, plus Alcoa (AA), American International Group (AIG), Bank of America (BAC), Citigroup (C), Hewlett-Packard (HPQ), Honeywell (HON), International Paper (IP), Altria Group (MO), and AT&T (T). Kraft (KRFT), General Motors (GM) and Kodak Eastman (KODK) are excluded due to trading disruptions and mergers, and Visa (V) is excluded due to having substantially less data available over the sample period (their initial public offering occurred in 2008).

All data are collected for the sample period of 2000-2014, inclusive. Equities markets were open for 3773 trading days in the sample period, whereas futures markets were open for 3790 days. In order to make all calculations consistent with the contemporaneous block bootstrap procedure, futures prices are dropped for the 17 days in which equities did not trade. Because the futures prices used are for the rolling front

month contract, there are occasional cases around expiry where no trading occurred. There were also a number of cases where trading did not occur due to lack of market interest. The majority of futures contracts had data for 3750 or more days, with exceptions being cocoa (3672 days, or 97.3%), cotton (3504/92.9%), coffee (3648/96.7%), orange juice (3245/86.0%) and sugar (3498/92.7%). In all calculations, missing data were not treated as generating zero returns, but rather were treated as explicitly generating missing returns, meaning that sample sizes for variance ratios and attendant bias corrections were adjusted accordingly.

4 Results

This section presents the results of our analysis. It begins by presenting the variance ratio testing results and showing the impact of the stepdown procedure on perceived statistical significance of the test statistics. Next, the test statistics are checked as to whether they conform to the assumptions of the stepdown procedure, and an explicit correction is applied as a robustness check. This adjustment results in nearly identical results. Next, results of the extended Turvey procedure are presented, and contrasted against the variance ratio testing results.

Tables 1 and 2 present the results of the estimation procedure detailed above for statistical significance of heteroskedasticity consistent variance ratios against a null hypothesis of 1.0. Table 1 presents results for futures contracts, while Table 2 presents results for equities, but results in both cases are adjusted for multiple hypothesis tests using the same stepdown procedure (i.e., the stepdown procedure generates both tables simultaneously). *, ** and *** denote statistical significance at the 95%, 99% and 99.9% levels respectively, and ^{a}, ^{b} and ^{c} denote 'significance' at the same levels without adjustment for the multiplicity of tests. All tests are two-sided. So, for example, $1.3236^{*,a}$ denotes a variance ratio of 1.3236 which is significantly different from 1.0 at the 95% level, with or without adjustment for multiple tests, while $0.9318^{*,b}$ denotes a variance ratio of 0.9318 that is significantly different from 1.0 at the 95% level, but which appears to be significant at the 99% level if the adjustment for multiple tests is ignored. When reading the table, note that variance ratios <1.0 (>1.0) correspond to Hurst parameters <0.5 (>0.5), indicating mean-reversion (momentum, or a trend-following property) in the underlying time series.

[Table 1 and Table 2 about here]

Among the highlights of Table 1 are that the feeder cattle (FC) contract at lags 2 and 5, and the oats (O) contract at lag 2, exhibit variance ratios rejecting independence at the 99% level, but which *appear* to reject independence at the 99.9% level if multiplicity of testing is not properly addressed. *In fact, no asset/lag combination tested achieves significance at the 99.9% level, or even the 99.7%* (3σ) *level recommended by Harvey et al. (2014)*. For equities, Table 2 shows that many asset/lag combinations achieve significance at the 95% level, but none do so at the 99% level, even though a large number *appear* to be significant at the 99% level and one stock, UTX (United Technologies Corp.), at lag 2 even *appears* to be significant at the 99.9% level when multiple testing is ignored. This pattern demonstrates that simple rules of thumb for disregarding marginal findings of statistical significance cannot replace a formalized

stepdown procedure accounting for the statistical dependence structure between all tests conducted, simultaneously. The role of heteroskedasticity is also important here. Referring back to Panel A of Figure 1, it is clear that many results of marginal statistical significance fall well outside the asymptotic confidence intervals, even when evaluating statistical significance of a single test.

As part of qualifying these results, it is important to verify that the test statistics generated in fact conform to the requirements of the stepdown procedure. The stepdown procedure of RW recommends studentized test statistics, but only requires that the statistics be reasonably well-behaved (see their paper for details). The heteroskedasticity consistent test statistic, φ^*, of LM is asymptotically unit normal, but not independent across equites or across time lags. The stepdown procedure exploits this non-independence, exhibited as a non-diagonal covariance matrix between the test statistics, to gain power while controlling FWER. However, there is no guarantee that the test statistics we generate are unit normal in finite samples, and in the case of this particular sample, some systematic violations are observed with respect to the lag, *k*.

Figure 2 shows how the first four standardized sample moments (mean, standard deviation, skewness, excess kurtosis) of the test statistics vary over lags in the bootstrapped sample ($I = 52$ assets, $T \approx 3770$ time periods, with 10,000 bootstrapped permutations of 1-period returns). The general pattern is that all the moments converge towards their asymptotic behavior over the short- to medium-term lags, but the skewness and excess kurtosis then begin to diverge away from zero for longer-term lags. It appears that the nature of Figure 2 arises because the moments of the distribution of test statistics converge towards their asymptotic levels at different rates according to the number of lags; other than the original derivation of asymptotic convergence, anecdotal evidence for this claim exists in the form of test statistics on the excluded Visa stock price data which have $T \approx 1700$ and produced a more exaggerated version of Figure 2.

[Figure 2 about here]

As a robustness check, we generate adjusted versions of the test statistics which were normalized by their bootstrapped mean and standard deviation to be in the form of *z*-scores: $\varphi' = \left(\varphi^* - \hat{\mu}_\varphi\right) / \hat{\sigma}_\varphi$. The goal of this procedure was to ensure that test statistics forced into the long tail of the distribution by their finite-sample properties (for specific lags) were compared on an equal scale to test statistics not exhibiting finite-sample departures from their asymptotic distribution. In Figure 3, Panel A shows the kernel densities of the LM bootstrapped test statistics, φ^*, by lag, while Panel B shows the same kernel densities after normalization into *z*-scores. Despite this adjustment (and despite the loss of aesthetic desirability in Panel A!), the results of the stepdown procedure were almost completely unchanged. In fact, the only change was to move oats (O), lag 2, from significance at the 99% level to significance at the 95% level. That said, it should be noted that the stepdown procedure does not admit precise *p*-values, only identification of whether a specific test statistic survives the stepdown procedure for a given FWER level. Regardless, the stability of the multiple testing results to lag-wise scaling lends substantial credibility to their robustness.

[Figure 3 about here]

Nonetheless, as discussed earlier, interpretation of the variance ratio testing results is difficult because a specific alternative hypothesis is not clearly identified. Instead, the variance ratio testing results *can* only show that there is some combined autocorrelation structure leading to violations of RWH at some lags, but not others. Worse yet, many failures to reject (even at the 95% level) are observed at large lags, and it cannot be known with certainty whether this is due to low statistical power for identifying these effects. Furthermore, there is the potential problem that variance ratios are affected by short-term dependence in the time series. For these reasons, the approach of Turvey to assess fractionality (or long-term memory) at the asset level may be preferable. We now present the estimation results of our algorithm extending Turvey, as outlined in Section 2 above.

Table 3 below shows the extended regression results with resulting estimates for the Hurst parameter, $\hat{H}$, at the asset level over the full time series. The table shows estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ estimated from Eq. 4 for both futures and equities, with statistical significance compared for individual hypothesis tests (^{a,b,c}) against explicit multiple testing (*,**,***), as in tables 1 and 2. It is immediately obvious in Table 3 that the constant term in the regression equation is significantly different from zero at the 99.9% level for every asset considered. While many of these values are quite small, some are larger than 0.1. The fBm-only model upon which Turvey's method is based does not explicitly consider the coefficient structure of the regression in Eq. 4, so it remains an area of future research to generalize the meaning of these estimates in the form of $\widehat{VR}(k) = k^{\alpha_0+2H}$. However, it is possible that these estimates are revealing a mixed-fBm model with a standard Brownian component, as discussed in Section 2.1 above and in Bender et al. (2007). It is also possible that the estimates are due to the combination of heteroskedasticity and non-normality of innovations in the observed data, but we leave these extensions and associated model checking as an area for future research.

[Table 3 about here]

It is also immediately obvious in the table that $\hat{\beta}_1$ is not significantly different from 1.0 for any of the futures contracts tested, though wheat *appears* to be significant at the 99% level when its hypothesis test is considered in isolation. For equities on the other hand, eleven of the 40 assets tested exhibit significant fractionality at the 99.9% level after controlling for multiple testing. Interestingly, every one of these exhibits an estimated Hurst parameter less than 0.5, implying a mean-reverting process, on average, as opposed to the (not significant) estimates for futures which vary above and below 0.5 about equally. We also note that the significance of the estimates of $\hat{\beta}_1$ and the corresponding values for $\hat{H}$ are consistent with both the pattern of estimated variance ratios in Table 2, and with the discussion in Section 2 indicating that weakening of variance ratio testing power over increasing lags could lead to failures to reject independence even when variance ratio patterns are strongly consistent with fBm. Table 3 also highlights the important role of controlling for heterogeneity in our extension of

Turvey's method: there are numerous cases where assets without significant fractionality have estimated values of $\hat{H}$ that are *further* from 0.5.

5 Discussion

It is clear from the evidence presented that variance ratio testing and regression estimation of the Hurst parameter are complementary approaches in attempting to understand the nature of financial time series. Variance ratio testing is best at identifying departures from the independence null for shorter time lags, while Hurst regression estimation is better at synthesizing the behavior of a time series into a test against a consolidated alternative hypothesis. It is possible that variance ratio tests have some substitutability with the constant term in the regression approach, but further modeling and empirical work is needed in this area, especially given the potential for mixed-fBm asset pricing models as an alternative explanation.

As discussed above, both approaches can give false rejections of the random walk hypothesis as a result of failing to explicitly adjust for heteroskedasticity and non-normality of innovations, so the new procedure presented here represents a step forward in consolidated testing of time series behavior. Furthermore, financial time series are necessarily evaluated both in the context of other time series simultaneously by the same researcher, and under the shadow of hypothesis testing by many researchers checking the same data over time. For these reasons, applied financial researchers will need to adjust findings to explicitly achieve an acceptable familywise error rate for the number of hypothesis tests considered, and this acceptable error rate may be decreasing over time as additional research is produced (Harvey et al., 2014). In light of this, the methods considered herein take these steps explicitly, and only results significant at the 3σ level should be considered as new discoveries in the area.

According to these standards, the results may be summarized as follows. For the 12 futures contracts and 40 U.S. equities considered over eight different time lags, none was discovered to violate time independence at the highest significance level after adjustment for the multiplicity of tests. This is evidence in support of the findings and assertions of Fama (1998), Turvey (2007), and Irwin (2013). On the other hand, when testing against a consolidated alternative hypothesis at the asset level, eleven equities were discovered to exhibit fractional, mean-reverting properties at the highest (multiplicity adjusted) 3σ significance level, consistent with Lo and Mackinlay's (1988) original results. Future research is needed to determine whether these results will hold when the analysis extends to the full set of listed equities and futures contracts, and to determine whether the variance ratio vs. fractional modeling approach is best. Future work will also determine the role of regression results in identifying mixed-fBm models (Bender et al., 2013) and the extent to which research designs in finance need to adapt explicit accounting for large numbers of hypothesis tests.

References

Acharya, V. V., Lochstoer, L. A., & Ramadorai, T. (2013). Limits to arbitrage and hedging: Evidence from commodity markets. *Journal of Financial Economics*, *109*(2), 441-465.

Asness, C. S., Moskowitz, T. J., & Pedersen, L. H. (2013). Value and momentum everywhere. *The Journal of Finance*, *68*(3), 929-985.

Bender, C., Sottinen, T. & Valkeila, E. (2007). Arbitrage with Fractional Brownian Motion? *Theory of Stochastic Processes,* 13(29), 23-34.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4), 1165-1188.

Bessembinder, H. (1992). Systematic risk, hedging pressure, and risk premiums in futures markets. *Review of Financial Studies*, *5*(4), 637-667.

Bollen, N. P., & Whaley, R. E. (2004). Does Net Buying Pressure Affect the Shape of Implied Volatility Functions? *The Journal of Finance*, *59*(2), 711-753.

Brunetti, C., & Reiffen, D. (2014). Commodity index trading and hedging costs. *Journal of Financial Markets*, *21*, 153-180.

Cao, C., Chen, Y., Liang, B., & Lo, A. W. (2013). Can hedge funds time market liquidity?. *Journal of Financial Economics*, *109*(2), 493-516.

Constantinides, G. M., Jackwerth, J. C., & Perrakis, S. (2009). Mispricing of S&P 500 index options. *Review of Financial Studies*, *22*(3), 1247-1277.

De Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*(3), 793-805.

Duffie, D. (2010). Presidential Address: Asset Price Dynamics with Slow-Moving Capital. *The Journal of Finance*, *65*(4), 1237-1267.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics*, *49*(3), 283-306.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work*. *The Journal of Finance*, *25*(2), 383-417.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 393-408.

Harvey, C. R., Liu, Y., & Zhu, H. (2014). ... *And the cross-section of expected returns* (No. w20592). National Bureau of Economic Research, Cambridge, MA.

Hirshleifer, D. (1990). Hedging Pressure and Futures Price Movements in a General Equilibrium Model. *Econometrica*, *58*(2), 411-28.

Hirshleifer, D. (1989). Determinants of hedging and risk premia in commodity futures markets. *Journal of Financial and Quantitative Analysis*, *24*(3), 313-331.

Hong, H., Kubik, J. D., & Fishman, T. (2012). Do arbitrageurs amplify economic shocks?. *Journal of Financial Economics*, *103*(3), 454-470.

Irwin, S. H. (2013). Commodity index investment and food prices: does the "Masters Hypothesis" explain recent price spikes? *Agricultural Economics*, *44*(s1), 29-41.

Irwin, S. H., & Sanders, D. R. (2011). Index Funds, Financialization, and Commodity Futures Markets. *Applied Economic Perspectives & Policy*, *33*(1), 1-31.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer Science & Business Media, Berlin.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data* (Vol. 53). John Wiley & Sons Incorporated, Hoboken.

Lo, A. W. (1991). LONG-TERM MEMORY IN STOCK MARKET PRICES. *Econometrica*, *59*(5), 1279-1313.

Lo, A. W., & MacKinlay, A. C. (1989). The size and power of the variance ratio test in finite samples: A Monte Carlo investigation. *Journal of Econometrics*, *40*(2), 203-238.

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies*, *1*(1), 41-66.

Mandelbrot, B. B., & Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM review*, *10*(4), 422-437.

Masters, M.W. (2008). Testimony before the Committee on Homeland Security and Government Affairs, U.S. Senate. Accessed December 2015. Available at http://www.hsgac.senate.gov//imo/media/doc/052008Masters.pdf

Romano, J. P., & Wolf, M. (2005a). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237-1282.

Romano, J. P., & Wolf, M. (2005b). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, *100*(469), 94-108.

Shiller, R. J. (1981). Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? *The American Economic Review*, *71*(3), 421-436.

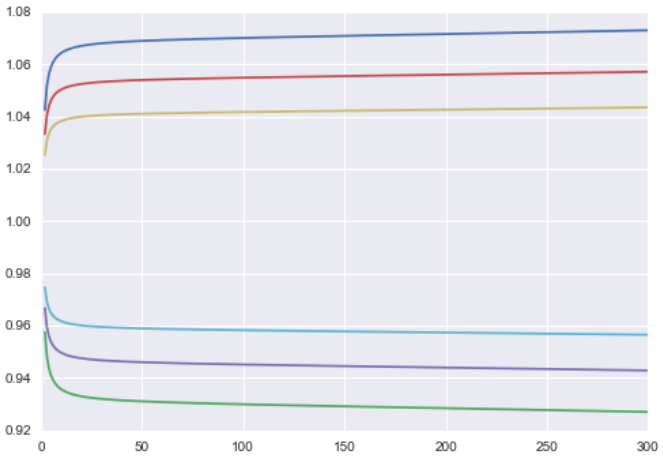
Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35-55.

Turvey, C. G. (2007). A note on scaled variance ratio estimation of the Hurst exponent with application to agricultural commodity prices. *Physica A: Statistical Mechanics and its Applications*, *377*(1), 155-165.

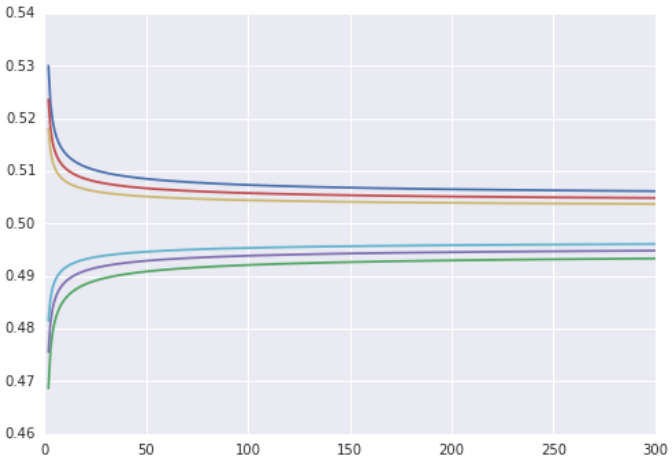
White, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press, Orlando, FL.

Figures

Figure 1: 95%, 99% and 99.9% Confidence Intervals for H_0 *(panels side-by-side if possible)*



Panel A: $H_0 : \widehat{VR}_1 = 1$



Panel B: $H_0 : \hat{H} = 0.5$

Figure 2: Bootstrapped Moments of φ^* by Lag

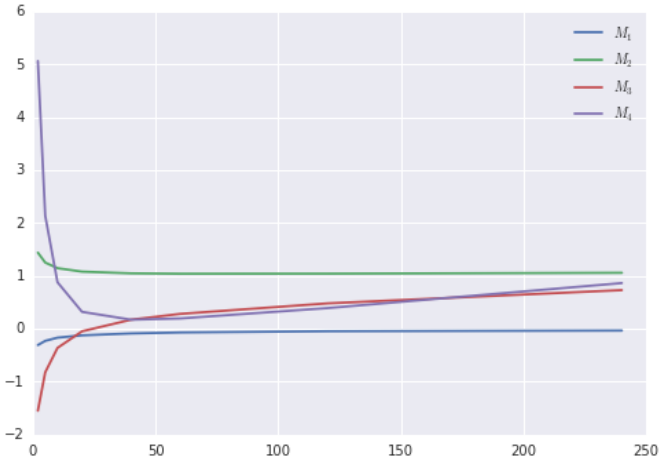
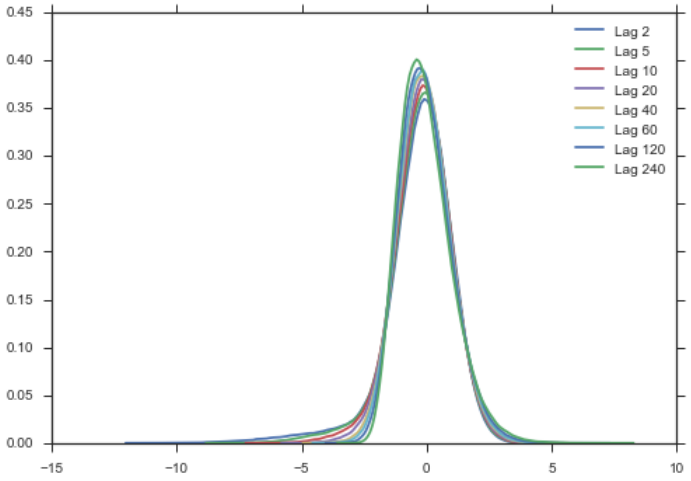
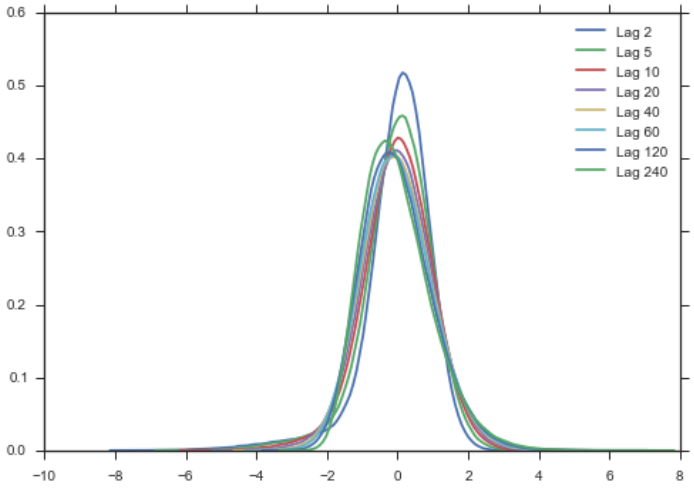


Figure 3: Kernel Densities of Test Statistics by Lag (panels side-by-side if possible)



Panel A: Finite Sample φ^*



Panel B: Adjusted φ'