Abstract

We perform a variety of white noise tests on daily stock market log returns over rolling windows of subsamples. Tests of weak form efficiency in stock returns are performed predominantly under the null hypothesis of independence or a martingale difference property. These properties rule out higher forms of dependence that may exist in stock returns that are otherwise serially uncorrelated. It is therefore of interest to test whether returns are white noise, allowing for a wider range of conditionally heteroskedastic time series, but also for non-martingale difference white noise. Assisted by the dependent wild bootstrap (Shao, 2010, 2011), we use sup-Lagrange Multiplier, Cramér-von Mises, and max-correlation statistics in order to test the white noise hypothesis. Evidently the dependent wild bootstrap has only been used in a full data sample, hence a key shortcoming has gone unnoticed: in rolling window sub-samples, the block structure unintentionally inscribes an artificial periodicity in computed p-values or confidence bands. We eliminate periodicity by randomizing the block size across bootstrap samples and windows. We find that the degree of market efficiency varies across countries and sample periods. In the case of Chinese and Japanese markets we cannot reject the white noise hypothesis, suggesting a high degree of efficiency. The same goes for markets in the U.K. and the U.S., provided trading occurs during non-crisis periods. When U.K. and U.S. markets face greater uncertainty, we tend to observe negative autocorrelations that are large enough to reject the white noise hypothesis.

JEL classifications: C12, C58, G14.
Keywords: dependent wild bootstrap, randomized block size, serial correlation, weak form efficiency, white noise test, maximum correlation test.

---

*We thank Yoshihiko Nishiyama and participants at the Econometric Seminar of the Institute of Economic Research, Kyoto University, 2016 Summer Workshop of Economic Theory, and 2016 Asian Meeting of the Econometric Society.
†Department of Economics, University of North Carolina, Chapel Hill. E-mail: jbhill@email.unc.edu
‡Graduate School of Economics, Kobe University. E-mail: motegi@econ.kobe-u.ac.jp
1 Introduction

We perform a variety of tests of weak market efficiency over rolling data sample windows, and make new contributions to the study of white noise tests. A stock market is \emph{weak form efficient} if stock prices fully reflect historical price information (Fama, 1970). Empirical results have been mixed, with substantial debate between advocates of the efficient market hypothesis (EMH) and proponents of behavioral finance.\footnote{See Yen and Lee (2008) and Lim and Brooks (2009) for an extensive survey of stock market efficiency.} More recently, the adaptive market hypothesis (AMH) proposed by Lo (2004, 2005) attempts to reconcile the two opposing schools, arguing that the degree of stock market efficiency varies over time.

In line with these trends, recent applications perform either non-overlapping subsample analysis or rolling window analysis in order to investigate the dynamic evolution of stock market efficiency. See Cheong, Nor, and Isa (2007), Hoque, Kim, and Pyun (2007), Lim (2008), and Urquhart and Hudson (2013) for use of non-overlapping subsamples, and for rolling windows see Kim and Shamsuddin (2008), Lim, Brooks, and Kim (2008), Kim, Shamsuddin, and Lim (2011), Lim, Luo, and Kim (2013), Verheyden, De Moor, and Van den Bossche (2015), Anagnostidis, Varsakelis, and Emmanouilides (2016), and Urquhart and McGroarty (2016). An advantage of rolling windows is that it does not require a subjective choice of the first and last dates of major events, including financial crises.

A common perception in the recent literature is that a financial crisis augments investor panic and hence lowers market efficiency. Cheong, Nor, and Isa (2007), Lim (2008), Lim, Brooks, and Kim (2008), and Anagnostidis, Varsakelis, and Emmanouilides (2016) find that market efficiency is indeed adversely affected by financial crises. By comparison, Hoque, Kim, and Pyun (2007), Kim and Shamsuddin (2008), and Verheyden, De Moor, and Van den Bossche (2015) find relatively mixed results.


Note that the implicit null hypothesis of all tests above is either that returns are iid, or a martingale difference sequence (mds) because the utilized asymptotic theory requires such assumptions under the null. These properties rule out higher forms of dependence that may exist in stock returns, while the mds property is generally not sufficient for a Gaussian central limit
theory (e.g. Billingsley, 1961). Chen and Deo (2006), for example, impose a martingale difference property on returns, and an eighth order unconditional cumulant condition. These are only shown to apply to GARCH and stochastic volatility processes with iid innovations, which ignores higher order dependence properties that arise under temporal aggregation (Drost and Nijman, 1993). Moreover, their test is not a true white-noise test since it does not test (asymptotically) that all serial correlations are zero.

A natural alternative is simply a white noise test with only serial uncorrelatedness under the null, as well as standard higher moment and weak dependence properties to push through standard asymptotics. A rejection of the white noise hypothesis might serve as a helpful signal for arbitragers, since a rejection indicates the existence of non-zero autocorrelation at some lags.

Formal white noise tests with little more than serial uncorrelatedness under the null have not been available until recently. See Hill and Motegi (2016a) for many detailed references, some of which are discussed below. Conventional portmanteau or Q-tests bound the maximum lag and therefore are not true white noise tests, although weak dependence, automatic lag selection, and a pivotal structure irrespective of model filter are allowed (e.g. Romano and Thombs, 1996, Lobato, 2001, Lobato, Nankervis, and Savin, 2002, Horowitz, Lobato, Nankervis, and Savin, 2006, Escanciano and Lobato, 2009, Delgado and Velasco, 2011, Guerre, Guerre, and Lazarova, 2013, Zhu and Li, 2015, Zhang, 2016).

Hong (1996, 2001) standardizes a portmanteau statistic, allowing for an increasing number of serial correlations and standard asymptotics. See also Hong and Lee (2003).


Hong (1999) generalizes the definition of the spectral density to allow for a test of independence (see also Hong and Lee, 2003). This is a powerful tool, but often in financial engineering and portfolio analysis serial independence is too strong a benchmark for studying market performance and determining efficiency. We place the present study in the literature that is strongly interested in whether asset returns are white noise, a useful albeit weak measure of market efficiency.

Hill and Motegi (2016a) develop a new theory for the maximum correlation test over an

---

increasing maximum lag. They allow for a very broad class of dependent and heterogeneous data, and verify that Shao’s (2011) dependent wild bootstrap is valid in this general setting. They compare the relative performance of multiple test statistics, using the dependent wild bootstrap to ensure correctly sized tests asymptotically. They find that Andrews and Ploberger’s (1996) sup-LM statistics and the Cramér-von Mises statistics used by Shao (2011) control for size fairly well. The max-correlation test, however, results in the sharpest empirical size according to their broad study. Overall, the max-correlation test yields either the highest power, or competitive power, and does not experience as sharp a decline in power when testing at very large lags. Indeed, the max-correlation test is specifically designed to be sensitive to dependence at large displacements, and relative to the above tests is best at detecting distant non-zero autocorrelations.

This paper uses Shao’s (2011) Cramér-von Mises test, Andrews and Ploberger’s (1996) sup-LM test, and Hill and Motegi’s (2016a) max-correlation test, assisted by the dependent wild bootstrap, in order to test whether stock returns are white noise. This analysis can be interpreted as testing whether stock prices follow what Campbell, Lo, and MacKinlay (1997) call Random Walk 3 (i.e. random walk with uncorrelated increment) in a strict sense.

We analyze daily stock price indices from China, Japan, the U.K., and the U.S. The entire sample period spans January 2003 through October 2015. As a preliminary analysis, we implement white noise tests for the full sample period. We then perform a rolling window analysis in order to capture subsample non-stationarity and therefore time-varying market efficiency. The degree of stock market efficiency may well be time-dependent given the empirical evidence from the previous literature on the adaptive market hypothesis. It is of particular interest to see how market efficiency is affected by financial turbulence like the subprime mortgage crisis around 2008.

We are not aware of any applications of the dependent wild bootstrap, except for Shao (2010, 2011) who analyzes temperature data and stock returns in a full sample framework. The present study is therefore the first use of the dependent wild bootstrap in a rolling window framework, in which we found and corrected a key shortcoming. In rolling window sub-samples, the block structure inscribes an artificial periodicity in the bootstrapped data, and therefore in computed p-values or confidence bands. A similar periodicity occurs in the block bootstrap for dependent data, which Politis and Romano (1994) correct by randomizing block size. We take the same approach to eliminate dependent wild bootstrap periodicity. See Section 2 for key details, and see the supplemental material Hill and Motegi (2016b) for complete details.

We find that the degree of market efficiency varies across countries and sample periods. Chinese and Japanese markets exhibit a high degree of efficiency since we generally cannot reject the white noise hypothesis. The same goes for the U.K. and the U.S. during non-crisis periods. When these
markets face greater uncertainty, for example during the Iraq War and the subprime mortgage crisis, we tend to observe negative autocorrelations that are large enough to reject the white noise hypothesis. A negative correlation, in particular at low lags, signifies rapid changes in market trading, which is corroborated with high volatility during these times. The appearance of negative autocorrelations in short (e.g. daily, weekly) and long (e.g. 1, 3 or 5-year) horizon returns has been documented extensively. Evidence for positive or negative correlations depends heavily on the market, return horizon (daily, weekly, etc.) and the presence of crisis periods. See, e.g., Fama and French (1988), who argue that predictable price variation due to mean-reversion in returns accounts for the negative correlation at short and long horizons.

The remainder of the paper is organized as follows. In Section 2 we explain the white noise tests that we use, and Section 3 describes our data. We report the empirical results for full sample and rolling sample analyses in Section 4. Concluding remarks are provided in Section 5.

2 Methodology

Let \( P_t \) be the stock market index at day \( t \in \{1, 2, \ldots, n\} \), and \( r_t = \ln(P_t/P_{t-1}) \) is the log-return. We assume returns are stationary in order to ensure that the various tests used this paper all have their intended asymptotic properties.\(^4\) Define the mean return \( \mu = E[r_t] \), and autocovariances \( \gamma(h) = E[(r_t - \mu)(r_{t-h} - \mu)] \) and autocorrelations \( \rho(h) = \gamma(h)/\gamma(0) \) for \( h \geq 0 \). We wish to test weak form efficiency:

\[
H_0 : \rho(h) = 0 \quad \text{for all } h \geq 1 \text{ against } H_1 : \rho(h) \neq 0 \text{ for some } h \geq 1.
\]

Similarly, write the sample mean return \( \hat{\mu}_n = 1/n \sum_{t=1}^{n} r_t \), autocovariance \( \hat{\gamma}_n(h) = 1/n \sum_{t=h+1}^{n} (r_t - \hat{\mu}_n)(r_{t-h} - \hat{\mu}_n) \) and autocorrelation \( \hat{\rho}_n(h) = \hat{\gamma}_n(h)/\hat{\gamma}_n(0) \) for \( h \geq 0 \). In order to ensure a valid white noise test and therefore capture all serial correlations asymptotically, we formulate test statistics based on the serial correlation sequence \( \{\hat{\rho}_n(h)\}_{h=1}^{L_n} \) with sample-size dependent lag length \( L_n \to \infty \) as \( n \to \infty \). We use tests by Andrews and Ploberger (1996), Shao (2011), and Hill and Motegi (2016a) due to their comparable size and power (cf. Hill and Motegi, 2016a).

The first of the three tests is the sup-LM test proposed by Andrews and Ploberger (1996).

\(^4\) We performed the Phillips and Perron (1988) test on market levels and differences: we fail to reject the unit root null hypothesis at any conventional level for levels, and reject the unit root null at the 1% level for differences. We performed three tests in each case: without a constant, with a constant and with a constant and linear time trend. The test statistics require a nonparametric variance estimator. We use a Bartlett kernel variance estimator with Newey and West’s (1994) automatic lag selection. P-values are computed using MacKinnon’s (1996) (one-sided) p-values.
The test statistic has the equivalent representation (see Nankervis and Savin, 2010):

\[ \mathcal{A} P_n = \sup_{\lambda \in \Lambda} \left\{ n(1 - \lambda^2) \left( \sum_{h=1}^{L_n} \lambda^{h-1} \hat{\rho}_n(h) \right)^2 \right\} \text{ where } L_n = n - 1, \]

where \( \Lambda \) is a compact subset of \((-1, 1)\). The latter ensures a non-degenerate test that obtains, under suitable regularity conditions, an asymptotic power of one when there is serial correlation at some horizon.

Andrews and Ploberger (1996) use \( L_n = n - 1 \) for computing the test statistic, but truncate a Gaussian series that arises in the limit distribution in order to simulate critical values. Nankervis and Savin (2010, 2012) generalize the sup-LM test to account for data dependence, and truncate the maximum lag both during computation (hence \( L_n < n - 1 \)), and for the sake of simulating critical values. The truncated value used, however, does not satisfy \( L_n \to \infty \) as \( n \to \infty \), hence their version of the test is not consistent (it does not achieve a power of one asymptotically when the null is false). To control for possible dependence under the null, and allow for a better approximation of the small sample distribution, we bootstrap the test with Shao’s (2011) dependent wild bootstrap, discussed below.

The second test is based on the following Cramér-von Mises [CvM] statistic used by Shao (2011):

\[ \mathcal{C}_n = n \int_0^{\pi} \left\{ \sum_{h=1}^{n-1} \hat{\gamma}_n(h) \psi_h(\lambda) \right\}^2 d\lambda \text{ where } \psi_h(\lambda) = (h\pi)^{-1} \sin(h\lambda). \]

By construction all \( n - 1 \) possible lags are used. The test statistic has a non-standard limit distribution under the null, and Shao (2011) demonstrates that a version of the dependent wild bootstrap proposed in Shao (2010) is valid under certain conditions on moments and dependence.

Third, the bootstrap max-correlation test proposed by Hill and Motegi (2016a) is based on the test statistic:

\[ \hat{T}_n = \sqrt{n} \max_{1 \leq h \leq L_n} |\hat{\rho}_n(h)|. \]

In this case \( L_n/n \to 0 \) is required such that \( \hat{\rho}_n(h) \) is Fisher consistent for \( \rho(h) \) for each \( 1 \leq h \leq L_n \). If the sequence of serial correlations were asymptotically iid Gaussian under the null then the limit law of a suitably normalized \( \hat{T}_n \) under the null is a Type I extreme value distribution, or Gumbel. That result extends to dependent data under the null (see Xiao and Wu, 2014, for theory and references). The non-standard limit law can be bootstrapped, as in Xiao and Wu (2014), although they do not prove their double blocks-of-blocks bootstrap is valid asymptotically. Hill and Motegi (2016a) sidestep an extreme value theoretic argument, and directly prove Shao’s (2011) dependent
wild bootstrap is valid without requiring the null limit law of the max-correlation.

Hill and Motegi (2016a) find that the above three tests have comparable size and power in finite samples, although the bootstrap max-correlation test yields the sharpest size in general, and in many cases obtains the highest power. Further, the max-correlation test is particularly sensitive to non-zero correlations at distant lags.\footnote{Other well-known test statistics for testing the white noise hypothesis include a standardized periodogram statistic of Hong (1996), which is effectively a standardized portmanteau statistic with a maximum lag $L_n = n - 1$. The test statistic has a standard normal limit under the null, but Hill and Motegi (2016a) show that an asymptotic test yields large size distortions. They also show that a bootstrap version, which is arithmetically equivalent to a bootstrapped portmanteau test, is often too conservative relative to the tests used in this study. We therefore do not include this test here.}

Each of the above three test statistics has a non-standard limit distribution under the null. We therefore use Shao’s (2011) dependent wild bootstrap in order to perform each test. Shao (2011) proves that the CvM test with the dependent wild bootstrap is asymptotically valid for processes that may exhibit a wide range of dependence under the null. Hill and Motegi (2016a) show that the dependent wild bootstrap is valid for the max-correlation test under an even larger class of dependent processes. The types of dependence allowed include various non-linear ARMA-GARCH processes with geometric (fast) or hyperbolic (slow) memory decay, and cover processes with or without a mixing property. Shao’s (2011) environment is only known to hold for geometric memory decay (see Shao, 2011, Hill and Motegi, 2016a).

The dependent wild bootstrap for the max-correlation test is executed as follows (sup-LM and CvM tests follow similarly). Set a block size $b_n$ such that $1 \leq b_n < n$. Generate iid random numbers $\{\xi_1, \ldots, \xi_{n/b_n}\}$ with $E[\xi_i] = 0$, $E[\xi_i^2] = 1$, and $E[\xi_i^4] < \infty$. Assume for simplicity that the number of blocks $n/b_n$ is an integer. Standard normal $\xi_i$ satisfies these properties, and is used in the empirical application below. Define an auxiliary variable $\omega_t$ block-wise as follows: $\{\omega_1, \ldots, \omega_{b_n}\} = \xi_1$, $\{\omega_{b_n+1}, \ldots, \omega_{2b_n}\} = \xi_2$, \ldots, $\{\omega_{(n/b_n-1)b_n+1}, \ldots, \omega_n\} = \xi_{n/b_n}$. Thus, $\omega_t$ is iid across blocks, but perfectly dependent within blocks. Compute:

$$\hat{\rho}^{(dw)}_n(h) = \frac{1}{\hat{\gamma}_n(0)} \frac{1}{n} \sum_{t=h+1}^{n} \omega_t \{ (r_t - \hat{\mu}_n)(r_{t-h} - \hat{\mu}_n) - \hat{\gamma}_n(h) \} \text{ for } h = 1, \ldots, L_n, \quad (1)$$

and a bootstrapped test statistic $\hat{T}^{(dw)}_n = \sqrt{n} \max_{1 \leq h \leq L_n} |\hat{\rho}^{(dw)}_n(h)|$. Repeat $M$ times, resulting in $\{\hat{T}^{(dw)}_{n,i}\}_{i=1}^M$. The approximate p-value is $\hat{p}^{(dw)}_{n,M} \equiv (1/M) \sum_{i=1}^{M} I(\hat{T}^{(dw)}_{n,i} \geq \hat{T}_n)$. Now let the number of bootstrap samples satisfy $M = M_n \to \infty$ as $n \to \infty$. If $\hat{p}^{(dw)}_{n,M} < \alpha$, then we reject the null hypothesis of white noise at significance level $\alpha$. Otherwise we do not reject the null. In our empirical study we use $M_n = 5,000$. The dependent wild bootstrapped CvM test and max-correlation tests are asymptotically valid.
and consistent, for large classes of processes that may be dependent under the null: the asymptotic probability of rejection at level $\alpha$ is exactly $\alpha$, and the asymptotic probability of rejection is one if the series is not white noise.\footnote{Besides the dependent wild bootstrap, Zhu and Li’s (2015) block-wise random weighting bootstrap can be applied to the CvM test statistic. Hill and Motegi (2016a) find that both bootstrap procedures are comparable in terms of empirical size and power.} It is also straightforward to show that the bootstrapped sup-LM test is asymptotically valid when the maximum lag is fixed for a similarly large class of dependent processes. We are not aware of a result in the literature that proves validity when the maximum lag is $L_n \to \infty$. Andrews and Ploberger (1996) and Nankervis and Savin (2010) use a simulation method based on a fixed maximum lag $L$ in order to approximate an asymptotic critical value. Simulations in Hill and Motegi (2016a), however, demonstrate that the bootstrapped test works well with $L_n$ increasing with $n$.

In some applications it is of interest to test for $\rho(h) = 0$ for a specific $h$. In that case (1) can be used to construct a bootstrapped confidence band under $\rho(h) = 0$. Compute $\{\hat{\rho}^{(dw)}_{n,i}(h)\}_{i=1}^{M}$ and sort them as $\hat{\rho}^{(dw)}_{n,[1]}(h) \leq \hat{\rho}^{(dw)}_{n,[2]}(h) \leq \cdots \leq \hat{\rho}^{(dw)}_{n,[M]}(h)$. The 95% band for lag $h$ is then $[\hat{\rho}^{(dw)}_{n,[.025*M]}(h), \hat{\rho}^{(dw)}_{n,[.975*M]}(h)]$. Below we compute those bands with $h = 1$ in the full sample and rolling window frameworks.

Evidently the dependent wild bootstrap has not been studied in a rolling window environment. Using the same auxiliary variable $\omega_t$ throughout one window of size $b_n \to \infty$, with $b_n = o(n)$, is key toward allowing for general dependence under the null. Unfortunately, it is easily shown that in a rolling window setting, the result is a periodically fluctuating p-value or confidence band, irrespective of the true data generating process (e.g. periodic fluctuations arise even for iid data). Thus, the dependent wild bootstrap does not generate stationary bootstrap samples in rolling windows of stationary data: an artificial seasonality across windows is present.

The reason for the periodicity is that we have similar blocking structures every $b_n$ windows. Consider two windows that are apart from each other by $b_n$ windows. A block in one window is a scalar multiplication of a block in the other window, resulting in similar bootstrapped autocorrelations from the two windows. See the supplemental material Hill and Motegi (2016b) for complete details and additional simulations that demonstrate the problem and solution, discussed below.

We solve the problem by randomizing the block size for each bootstrap sample and window. Randomness across windows ensures that different windows have different blocking structures, and are therefore not multiples of each other. This removes the artificial nonstationarity, conditional on the sample. Randomness across bootstrap samples makes the confidence bands less volatile, which is desired in terms of visual inspection. In a full sample we find using values considered in Shao (2011), $b_n = c \sqrt{n}$ with $c \in \{1/2, 1, 2\}$, to be comparable, hence we simply use $c = 1$ for all
bootstrap samples and windows. In rolling windows we therefore draw a uniform random variable $c$ on $[0.5, 1.5]$ for each bootstrap sample and window, and use $b_n = c\sqrt{n}$.

Randomizing a bootstrap block size to ensure stationary bootstrap draws (conditional on the sample) is not new. Politis and Romano (1994) use randomized block sizes in their stationary bootstrap subsampling method, a block bootstrap procedure that ensures a stationary bootstrap draw. See Lahiri (1999) for enhanced theoretical properties. In a similar sense, randomizing block size for the wild bootstrap removes artificially introduced non-stationarity in the bootstrap draws over rolling windows, conditional on the sample.

3 Data

We analyze log returns of the daily closing values from the Chinese Shanghai Composite Index (Shanghai), the Japanese Nikkei 225 index (Nikkei), the U.K. FTSE 100 Index (FTSE), and the U.S. S&P 500 index (SP500), all in local currencies from January 1, 2003 through October 29, 2015. The Shanghai index is selected as a representative of emerging markets, while the latter three are known as some of the most liquid, mature, and influential markets. The sample size differs across countries due to different trading days, holidays, and other market closures: 3110 days for Shanghai, 3149 days for Nikkei, 3243 for FTSE, and 3230 for SP500. Market closures are simply ignored, hence the sequence of returns are treated as daily for each observation.

Figure 1 plots each stock price index and the log return. The subprime mortgage crisis in 2007-2008 caused a dramatic decline in the stock prices. Comparing the first and last trading days in 2008, the growth rate of stock price level is -65.5% for Shanghai, -39.7% for Nikkei, -30.9% for FTSE, and -37.6% for SP500. The latter two countries experienced relatively fast recovery from the stock price plummet in 2008, while Shanghai and Nikkei experienced a longer period of stagnation. Each return series shows clear volatility clustering, especially during the crisis.

Besides the subprime mortgage crisis, there are a few episodes of financial turbulence that may affect the autocorrelation structure of each market. First, in 2014-2015, the Shanghai stock market experienced an apparent “bubble” (and collapse) the magnitude of which is only slightly smaller than the subprime mortgage crisis. In the first half of 2014, the Shanghai index fluctuated steadily around 2050 points. It then soared for a year to the peak of 5166.35 on June 12, 2015, and then fell down sharply to 2927.29 on August 26, 2015.

Second, Nikkei experienced a large negative log return of -0.112 on March 15, 2011, which is two business days after the Great East Japan Earthquake. Nuclear power plants in Fukushima were destroyed by a resulting tsunami, and there emerged pessimistic sentiment among investors.
on electricity supply.

Third, in 2002 and 2003, the FTSE stock market faced a period of great uncertainty due to an economic recession, soaring oil prices, and the Iraq War. FTSE fell for nine consecutive trading days in January 2003 losing 12.4% of its value. On March 13, 2003 the FTSE experiences a rebound gain of .059, highlighting an unstable market condition.

Fourth, the SP500 index generated a log return of -0.069 on August 8, 2011 because Standard & Poor’s downgraded the SP500 federal government credit rating from AAA to AA+ on August 5.

Insert Figure 1 here

Table 1 lists sample statistics of return series. Each series has a positive mean, but it is not significant at the 5% level according to a bootstrapped confidence band. Shanghai returns have the largest standard deviation, but Nikkei returns have the greatest range: it has the largest minimum and maximum in absolute value. Each series displays negative skewness and has a large kurtosis, all stylized traits. The kurtoses are 6.831 for Shanghai, 10.68 for Nikkei, 11.14 for FTSE, and 14.02 for SP500. Shanghai has much smaller kurtosis than the other countries, but it is still much larger than the kurtosis of the normal distribution. Due to the negative skewness and excess kurtosis, the p-values of the Kolmogorov-Smirnov and Anderson-Darling tests of normality are well below 1% for all countries, strong evidence against normality.

Insert Table 1 here

4 Empirical Results

We now present the main empirical findings.

4.1 Full Sample Analysis

Figure 2 shows sample autocorrelations of the daily return series from January 1, 2003 through October 29, 2015. Lags $h = 1, \ldots, 25$ trading days are considered. The 95% confidence bands are constructed with Shao’s (2011) dependent wild bootstrap under the null hypothesis of white noise. Hence a sample autocorrelation lying outside the confidence band can be thought of as an evidence against the white noise hypothesis, conditional on each lag $h$. The number of bootstrap samples is 5,000, and in all cases we draw from the standard normal distribution.

Insert Figure 2 here
While most Shanghai sample correlations lie inside the 95% bands, there are some marginal cases. Correlations at lags 3 and 4, for example, are respectively 0.036 and 0.064, which are on or near the upper bounds. Further, Nikkei, FTSE, and SP500 have significantly negative correlations at lag 1. The sample correlation with accompanied confidence band is -0.036 with [-0.031, 0.031] for Nikkei; -0.054 with [-0.042, 0.042] for FTSE; -0.102 with [-0.079, 0.075] for SP500. Thus, returns on each of these indices are unlikely to be white noise. Negative correlations at lag 1 suggest that traders switch between short and long positions actively on a daily basis, and as a result a sharp price hike tends to be followed by a sharp price drop and vice versa.

In general, we observe insignificant autocorrelations at large lags (e.g. $h > 15$) for each series, which suggests that there does not exist a noticeable seasonal effect.8

Table 2 compiles bootstrapped p-values from the max-correlation, sup-LM, and CvM white noise tests. The max-correlation test requires the maximum lag length $L_n = o(n)$, while the sup-LM and CvM tests use $L_n = n - 1$. Nankervis and Savin (2010) truncate the correlation series in the sup-LM statistic, using $L_n = 20$ for each $n$, which fails to deliver a consistent test. We use each $L_n = \max\{5, [\delta \times n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$ for the max-correlation and sup-LM tests for comparability, as well as $L_n = n - 1$ for sup-LM and CvM tests. Note that, under suitable regularity conditions, as long as $L_n \to \infty$ then the sup-LM and CvM tests will have their intended limit properties under the null and alternative hypotheses, even if $L_n = o(n)$. In the case of Shanghai, for example, we have $n = 3110$ and hence $L_n \in \{5, 77, 154, 193, 386\}$. The bootstrap block size is set to be $b_n = [\sqrt{n}]$ as in Hill and Motegi (2016a), but values like $b_n = [(1/2)\sqrt{n}]$ or $b_n = [2\sqrt{n}]$ lead to similar results, cf. Shao (2011).

The strongest evidence for correlation in Shanghai comes from the max-correlation test at maximum lag 5. As more lags are incorporated into the max-correlation and sup-LM statistics, the bootstrapped p-values are progressively larger, and the CvM test with all possible lags likewise fails to reject the null hypothesis. The evidence therefore favors a rejection of weak form efficiency in the Shanghai market.

Contrary to the case of Shanghai, for Nikkei, FTSE, and SP500 the max-correlation test leads to a non-rejection and the sup-LM and CvM tests lead to a rejection. The max-correlation p-value is exactly .10 at lag 5 for SP500, suggesting weak evidence against white noise. There are several explanations. First, this may be because Nikkei, FTSE, and S&P 500 have barely significant (and small) negative autocorrelations at lag 1, as we saw in Figure 2, with significance at the 10% level for Nikkei and FTSE, and 5% for the SP500. Second, the max-correlation test has the sharpest

---

8 Results with larger lags are available upon request.
empirical size amongst these tests, while sup-LM and CvM tests tend to be over-sized. Thus, we may be viewing false rejections from sup-LM and CvM tests.

Each of the four series, however, has fairly weak serial correlation over each lag, with any evidence of correlation coming at low lags. sup-LM and CvM tests result in greater sampling error and therefore lower power as lags are added, and tend to be over-sized even at low lags. Thus, if there is evidence against weak form efficiency in these markets it will generally come at low lags. With that in mind, the max-correlation test, with generally sharper size and greater power\(^9\), suggests the Shanghai index fails to be weak form efficient over the full sample, with evidence against efficiency in the SP500.

4.2 Rolling Window Analysis

A potential drawback of the full sample analysis is that the degree of stock market efficiency may not be constant throughout the sample period in reality, as pointed out by Lo (2004, 2005). We therefore perform a rolling window analysis in order to capture the dynamic nature of market efficiency. An advantage of this approach is that we can examine whether negative autocorrelations arising in Nikkei, FTSE, and SP500 is a persistent phenomenon. We set the window size to be 240 trading days (roughly a year), which is similar to the window size in Verheyden, De Moor, and Van den Bossche (2015).

4.2.1 Preliminary Analysis

As a preliminary analysis, Figure 3 plots the standard deviation, minimum, maximum, skewness, and kurtosis of each return series in each rolling window. As expected, the standard deviation for each country blows up around 2008, reflecting the subprime mortgage crisis. Shanghai’s standard deviation in 2015 is roughly as large as what it used to be in 2008, reflecting the Shanghai stock market collapse.

There is a dramatic change in skewness and kurtosis for Nikkei and SP500 in 2010-2011. Skewness suddenly reaches a large negative value of about -2, and kurtosis skyrockets to 20 for Nikkei and to 14 for the SP500. These results stem from the tsunami disaster and S&P securities downgrade shock explained in Section 3.

In early 2003, the FTSE had a large positive skewness of around 0.8 and a large kurtosis of around 7, reflecting the Iraq War and the recession. The positive skewness stems from a large log

---

\(^9\)Simulation experiments performed in Hill and Motegi (2016a) reveal that the dependent wild bootstrapped max-correlation test over most cases considered has comparatively sharper size and higher power than the sup-LM and CvM tests, as well as a variant of Hong’s (1996) tests and the bootstrapped Q-test.
return of 0.059 realized on March 13, 2003, when it was a bear market.

4.2.2 Analysis of Serial Correlation

Figure 4 presents first order sample autocorrelations over rolling windows. For each window, the 95% confidence band based on the dependent wild bootstrap is constructed under the null hypothesis of white noise. Window size is \( n = 240 \), and we begin with a conventional block size \( b_n = \lceil \sqrt{n} \rceil = 15 \). The number of bootstrap samples is 10,000 for each window.

A striking result from Figure 4 is that the confidence bands exhibit periodic fluctuations, with the appearance of veritable seasonal highs and lows. Moreover, the zigzag movement repeats itself in every \( b_n = 15 \) windows. The confidence bands are particularly volatile for Nikkei and SP500 in 2011, reflecting the tsunami disaster and S&P securities downgrade shock. Note, however, that periodic confidence bands appear in all series and periods universally.

In the supplemental material Hill and Motegi (2016b) we elaborate on this phenomenon, with computational details, and magnified plots for ease of viewing. As discussed in Section 2, the periodicity arises because there are similar blocking structures every \( b_n \) windows. Data blocks from two windows separated by \( b_n \) windows are scalar multiples of each other, resulting in similar bootstrapped autocorrelations from the two windows. A similar issue arises in samples generated by block bootstrap (see, e.g., Lahiri, 1999). The solution proposed by Politis and Romano (1994) is to randomize block size, which we follow. For each of the 10,000 bootstrap samples, and each window, we independently draw \( c \) from a uniform distribution on \([0.5, 1.5]\), and use \( b_n = c\sqrt{n} \).

A comparison of Figures 4 and 5 highlights the substantial impact of block size randomization: in Figure 5 randomization has clearly removed the periodic fluctuations.\(^{11}\) We still observe volatile bands for Nikkei right after the tsunami disaster and for SP500 right after the U.S. securities downgrade shock. These are not surprising results since the correlations themselves spike in those periods. In what follows, we focus the discussion on results based on block randomization, and

\(^{10}\)If we only randomize \( b_n \) for each window (using the same randomized \( b_n \) across all 10,000 bootstrap samples), then there exists volatility in the resulting confidence bands that largely exceeds the volatility of the observed data. By randomizing \( b_n \) for each bootstrap draw and each window, both artificial nonstationarity and excess volatility are eradicated.

\(^{11}\)See also the supplemental material for related plots from controlled experiments. In particular, we present magnified plots of bootstrapped confidence bands from simulated data, with and without randomized block size.
For Shanghai, the confidence bands are roughly \([-0.1, 0.1]\) and they contain the sample correlation in most windows. Interestingly, the subprime mortgage crisis around 2008 did not have a substantial impact on the correlation structure of the Shanghai market, although the stock price itself responded with a massive drop and volatility burst around 2008 (see Figures 1 and 3). In 2014-2015, the confidence bands are slightly wider, reflecting the Shanghai stock market collapse. The correlations sometimes go beyond 0.1 and outside the confidence bands. Hence, conditional on lag \(h = 1\), there is a possibility that the Shanghai collapse had an adverse impact on the market efficiency. This is clearly a richer implication than what we saw from the full sample analysis, highlighting an advantage of rolling window analysis.

The first order correlation for Nikkei generally lies in \([-0.1, 0.1]\) and they are insignificant in most windows. The correlation goes beyond 0.1 in only one out of 2910 windows, which is window \#1772 (March 24, 2010 - March 15, 2011). This is the first window that contains the tsunami shock. The correlation is 0.142 and the confidence band is \([-0.216, 0.208]\) there, hence the zero hypothesis is not rejected. In terms of negative correlations, we have \(\hat{\rho}_n(1) < -0.1\) in 210 windows (approximately 7.2% of all windows). The zero hypothesis is rejected for 25 out of the 210 windows. This result is consistent with the negative correlation at lag 1 observed in the full sample analysis.

The tendency for negative correlations is much more prominent in FTSE. The correlation for FTSE goes below -0.1 for 954 windows out of 3004. A rejection occurs in 478 out of the 954 windows. The correlation goes even below -0.2 for 106 windows, and a rejection occurs in 66 windows.

Similarly, SP500 often has significantly negative correlations. The correlation for SP500 goes below -0.1 for 1109 windows out of 2991. A rejection occurs in 675 out of the 1109 windows. The correlation goes even below -0.2 for 31 windows, and a rejection occurs in 18 windows.

It is interesting that the more mature and liquid markets have a stronger tendency for having negative correlations. Negative serial correlations may be evidence of mean reversion, and therefore long run stationarity (see Fama and French, 1988), which fits our maintained assumption that log-prices are first difference stationary.

Another implication from Figure 5 is that the significantly negative correlations concentrate on the period of financial turmoil: Iraq-War regime in 2003 for FTSE and SP500 and the subprime mortgage crisis in 2008 for SP500. In view of the extant evidence for positive serial correlation in market returns, this suggests that negative correlations may be indicative of trading turmoil, due ostensibly to the rapid evolution of information.
Evidence for negative correlations during crisis periods is not new. See, for example, Campbell, Grossman, and Wang (1993) who find a negative relationship between trading volume and serial correlation: high volume days are associated with lower or negative correlations, due, they argue, to the presence of "noninformational" traders.

4.2.3 White Noise Tests

We now perform the max-correlation, sup-LM, and CvM tests for each window. As in Section 4.1, lag length is $L_n = \max\{5, [\delta \times n / \ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$ for the max-correlation and sup-LM tests. Since the sample size is constant at window size $n = 240$, we have that $L_n = 5, 8, 17, 21, 43$. We also cover $L_n = n - 1 = 239$ for the sup-LM and CvM tests. Block size is $b_n = c\sqrt{n}$, and we draw $c$ from the uniform distribution between 0.5 and 1.5 independently across $M = 5,000$ bootstrap samples and rolling windows.

See Figures 6-8 for rolling window test statistics, critical values, and p-values as well as Table 3 for summary results. Our results suggest that the stock markets of Shanghai and Nikkei are likely weak form efficient throughout the whole sample period. Based on the max-correlation test with lag 5, for example, a rejection happens in only 3.7% of all windows for Shanghai and 3.3% for Nikkei. We observe similar results based on the sup-LM and CvM tests. Recall, however, that Shanghai has positive correlations that are barely significant at the 5% level in 2014-2015 (see Figure 5). These are only first order correlations, and their magnitude is not too large, while statistical significance of each tests declines when lags $L_n \geq 5$ are considered jointly.

Insert Figures 6-8 here

The FTSE and SP500 market have more periods of inefficiency than Shanghai and Nikkei. Based on the CvM test, for example, a rejection happens in as many as 14.5% of all windows for FTSE and 21.0% for SP500. The sup-LM test with any lag selection yields similar results, while the max-correlation test produces lower rejection frequencies (less than 10%). This difference is reasonable since the negative correlation at lag 1 should be better captured by the sup-LM and CvM tests than the max-correlation test (cf. simulation experiments in Hill and Motegi, 2016a,b). As seen in Figures 7-8, rejections occur continuously during the Iraq War and the subprime mortgage crisis. The former has a longer impact than the latter for FTSE, while the latter has a longer impact for SP500. This result is consistent with Figure 5. It is also consistent with the notion that high volatility is associated with lower or negative correlations, since these periods are marked by increased volatility.

Insert Table 3 here
In summary, the degree of stock market efficiency differs noticeably across countries and sample periods, as asserted by the adaptive market hypothesis. The Shanghai and Nikkei stock markets have a high degree of efficiency so that we cannot reject the white noise hypothesis of stock returns. The same goes for FTSE and SP500 during non-crisis periods. When they are in unstable periods like the Iraq War and the subprime mortgage crisis, we tend to observe negative autocorrelations that are large enough to reject the white noise hypothesis. Thus, conditional on being in a non-crisis period, the latter markets are efficient, but fare less well in terms of efficiency during crisis periods, relative to Asian markets.

5 Conclusion

Much of the previous literature on testing for weak form efficiency of stock markets imposes an iid or mds assumption under the null hypothesis. The iid property rules out any form of conditional heteroskedasticity, and the mds property rules out higher level forms of dependence. It is thus of interest to perform tests of the white noise hypothesis with little more than serial uncorrelatedness under the null hypothesis. A rejection of the white noise hypothesis might serve as a helpful signal of an arbitrage opportunity for investors, since it indicates the presence of non-zero autocorrelation at some lags.

Using theory developed in Hill and Motegi (2016a) that extends Shao’s (2011) dependent wild bootstrap to max-correlation and sup-LM tests, as well as Shao’s (2011) proposed Cramér-von Mises test, we analyze weak form efficiency of Chinese, Japanese, U.K., and U.S. stock markets.

We perform both full sample and rolling window analyses, where the latter allows us to capture time-varying market efficiency. The present study is apparently the first use of the dependent wild bootstrap in a rolling window environment. The block structure inscribes an artificial periodicity in computed p-values or confidence bands over rolling windows, which is removed by randomizing block sizes.

We find that the degree of market efficiency varies across countries and sample periods. The Shanghai and Nikkei returns exhibit a high degree of efficiency such that we generally cannot reject the white noise hypothesis. The same goes for the FTSE and SP500 during non-crisis periods. When those markets face greater uncertainty, for example during the Iraq War and the subprime mortgage crisis, we tend to observe negative autocorrelations that are large enough to reject the white noise hypothesis. A negative correlation, in particular at low lags, signifies rapid changes in market trading, which is corroborated with high volatility due to noninformational traders.
References


Table 1: Sample Statistics of Log Returns of Stock Price Indices (01/01/2003 - 10/29/2015)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Shanghai</td>
<td>3110</td>
<td>2.9 × 10^{-4}</td>
<td>[−7.4, 7.2] × 10^{-4}</td>
<td>0.001</td>
<td>0.017</td>
<td>-0.093</td>
<td>0.090</td>
<td>-0.425</td>
<td>6.831</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Nikkei</td>
<td>3149</td>
<td>2.5 × 10^{-4}</td>
<td>[−5.0, 5.1] × 10^{-4}</td>
<td>0.001</td>
<td>0.015</td>
<td>-0.121</td>
<td>0.132</td>
<td>-0.532</td>
<td>10.68</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>FTSE</td>
<td>3243</td>
<td>1.5 × 10^{-4}</td>
<td>[−2.7, 2.6] × 10^{-4}</td>
<td>0.001</td>
<td>0.012</td>
<td>-0.093</td>
<td>0.094</td>
<td>-0.133</td>
<td>11.14</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>SP500</td>
<td>3230</td>
<td>2.7 × 10^{-4}</td>
<td>[−3.8, 3.7] × 10^{-4}</td>
<td>0.001</td>
<td>0.012</td>
<td>-0.095</td>
<td>0.110</td>
<td>-0.319</td>
<td>14.02</td>
<td>0.000</td>
<td>0.001</td>
</tr>
</tbody>
</table>

"95% Band" is a bootstrapped 95% confidence band for the sample mean. It is constructed under the null hypothesis of zero-mean white noise, using the dependent wild bootstrap with block size \( b_n = \sqrt{n} \). The number of bootstrap samples is \( M = 10,000 \). "p-KS" signifies a p-value of the Kolmogorov-Smirnov test, while "p-AD" signifies a p-value of the Anderson-Darling test.
Table 2: P-Values of White Noise Tests over the Full Sample (01/01/2003 - 10/29/2015)

<table>
<thead>
<tr>
<th></th>
<th>Shanghai (3110 Days)</th>
<th>Nikkei (3149 Days)</th>
<th>FTSE (3243 Days)</th>
<th>SP500 (3230 Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max-Correlation</td>
<td>Andrews-Ploberger</td>
<td>CvM</td>
<td>Max-Correlation</td>
</tr>
<tr>
<td>( \delta = 0.0 )</td>
<td>( 0.053 )</td>
<td>( 0.053 )</td>
<td>( 0.094 )</td>
<td>( 0.094 )</td>
</tr>
<tr>
<td>( \delta = 0.2 )</td>
<td>( 0.314 )</td>
<td>( 0.314 )</td>
<td>( 0.317 )</td>
<td>( 0.317 )</td>
</tr>
<tr>
<td>( \delta = 0.4 )</td>
<td>( 0.398 )</td>
<td>( 0.398 )</td>
<td>( 0.427 )</td>
<td>( 0.427 )</td>
</tr>
<tr>
<td>( \delta = 0.5 )</td>
<td>( 0.447 )</td>
<td>( 0.447 )</td>
<td>( 0.563 )</td>
<td>( 0.563 )</td>
</tr>
<tr>
<td>( \delta = 1.0 )</td>
<td>( 0.182 )</td>
<td>( 0.182 )</td>
<td>( 0.425 )</td>
<td>( 0.425 )</td>
</tr>
<tr>
<td>( \mathcal{L}_n = 5 )</td>
<td>( 0.083 )</td>
<td>( 0.083 )</td>
<td>( 0.094 )</td>
<td>( 0.094 )</td>
</tr>
<tr>
<td>( \mathcal{L}_n = 77 )</td>
<td>( 0.065 )</td>
<td>( 0.065 )</td>
<td>( 0.162 )</td>
<td>( 0.162 )</td>
</tr>
<tr>
<td>( \mathcal{L}_n = 154 )</td>
<td>( 0.079 )</td>
<td>( 0.079 )</td>
<td>( 0.169 )</td>
<td>( 0.169 )</td>
</tr>
<tr>
<td>( \mathcal{L}_n = 193 )</td>
<td>( 0.074 )</td>
<td>( 0.074 )</td>
<td>( 0.177 )</td>
<td>( 0.177 )</td>
</tr>
<tr>
<td>( \mathcal{L}_n = 386 )</td>
<td>( 0.418 )</td>
<td>( 0.418 )</td>
<td>( 0.342 )</td>
<td>( 0.342 )</td>
</tr>
<tr>
<td>( \mathcal{L}_n = 3109 )</td>
<td>( 0.157 )</td>
<td>( 0.157 )</td>
<td>( 0.328 )</td>
<td>( 0.328 )</td>
</tr>
</tbody>
</table>

Bootstrapped p-values of Hill and Motegi’s (2016a) max-correlation white noise test, Andrews and Ploberger’s (1996) sup-LM test, and the Cramér-von Mises test. Shao’s (2011) dependent wild bootstrap with 5000 replications is used for each test. The maximum lag lengths for the max-correlation and sup-LM tests are \( \mathcal{L}_n = \max\{5, [\delta \times n/\ln(n)] \} \) with \( \delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\} \). The sup-LM test is also computed with the maximum possible \( n - 1 \) lags. The CvM test uses \( \mathcal{L}_n = n - 1 \).
Table 3: Rejection Ratio of White Noise Tests over Rolling Windows

<table>
<thead>
<tr>
<th></th>
<th>Max-Correlation</th>
<th>Andrews-Ploberger</th>
<th>CvM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\delta = 0.0$</td>
<td>$\delta = 0.2$</td>
<td>$\delta = 0.4$</td>
</tr>
<tr>
<td></td>
<td>$L_n = 5$</td>
<td>$L_n = 8$</td>
<td>$L_n = 17$</td>
</tr>
<tr>
<td>Shanghai</td>
<td>0.037</td>
<td>0.025</td>
<td>0.004</td>
</tr>
<tr>
<td>Nikkei</td>
<td>0.033</td>
<td>0.039</td>
<td>0.000</td>
</tr>
<tr>
<td>FTSE</td>
<td>0.072</td>
<td>0.043</td>
<td>0.016</td>
</tr>
<tr>
<td>SP500</td>
<td>0.088</td>
<td>0.031</td>
<td>0.016</td>
</tr>
</tbody>
</table>

The ratio of rolling windows where the null hypothesis of white noise is rejected at the 5% level. Test statistics are Hill and Motegi's (2016a) max-correlation statistic, Andrews and Ploberger’s (1996) sup-LM statistic, and the Cramér-von Mises statistic. We use Shao’s (2011) dependent wild bootstrap with block size $b_n = [c \times \sqrt{n}]$. We draw $c \sim U(0.5, 1.5)$ independently across $M = 5,000$ bootstrap samples and rolling windows. The maximum lag lengths for the max-correlation and sup-LM tests are $L_n = \max\{5, [\delta \times n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$ and window size $n = 240$ trading days. We also cover $L_n = n - 1$ for the sup-LM and CvM tests.
Left panels depict stock price indices in local currencies (standardized at 100 on 01/01/2003). Note that the maximum value of the vertical axis is 500 for Shanghai and 300 for the other series. Right panels depict log returns. "Subprime" signifies the subprime mortgage crisis around 2008; "Shanghai" signifies the 2014-2015 bubble and burst in Shanghai; "Tsunami" signifies the tsunami disaster caused by the Great East Japan Earthquake in March 2011; "Iraq" signifies Iraq War (plus economic stagnation) threatening the United Kingdom; "AA+" signifies the downgrade of the U.S. federal government credit rating from AAA to AA+. 

24
This figure plots sample autocorrelations at lags 1, . . . , 25. The 95% confidence bands are constructed with Shao’s (2011) dependent wild bootstrap under the null hypothesis of white noise. The number of bootstrap samples is 5,000.
Standard deviation (std), minimum (min), maximum (max), skewness (skew), and kurtosis (kurt) of stock returns of Shanghai, Nikkei, FTSE, and SP500 in each rolling window with size 240 trading days. Each point on the horizontal axis represents the initial date of each window.
The solid line depicts sample autocorrelations at lag 1. The dotted line depicts the 95% confidence band constructed with Shao’s (2011) dependent wild bootstrap under the null hypothesis of white noise. Window size is $n = 240$ trading days, and block size is $b_n = \lfloor \sqrt{n} \rfloor = 15$. The number of bootstrap samples is 10,000 for each window. Each point on the horizontal axis represents the initial date of each window.
Figure 5: Rolling Window Autocorrelations at Lag 1 (Randomized Block Size)

The solid line depicts sample autocorrelations at lag 1. The dotted line depicts the 95% confidence band constructed with Shao’s (2011) dependent wild bootstrap under the null hypothesis of white noise. Window size is $n = 240$ trading days. Block size is $b_n = \sqrt{n}$. We draw $c \sim U(0.5, 1.5)$ independently across $M = 10,000$ bootstrap samples and rolling windows. Each point on the horizontal axis represents the initial date of each window. “Shanghai” signifies the 2014-2015 bubble and burst in Shanghai; “Tsunami” signifies the tsunami disaster caused by the Great East Japan Earthquake in March 2011; “Iraq” signifies Iraq War (plus economic stagnation) threatening the United Kingdom; “Subprime” signifies the subprime mortgage crisis around 2008; “AA+” signifies the downgrade of the U.S. federal government credit rating from AAA to AA+. 
Hill and Motegi’s (2016a) max-correlation test. Lag length is $\mathcal{L}_n = \max\{5, \lceil \delta \times n / \ln(n) \rceil \}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$ and window size $n = 240$ trading days, hence $\mathcal{L}_n = 5, 8, 17, 21, 43$. We use the dependent wild bootstrap with 5,000 replications for each window. Block size is $b_n = \lceil c \sqrt{n} \rceil$, where $c \sim U(0.5, 1.5)$ is drawn independently across bootstrap samples and windows. Each point on the horizontal axis represents the initial date of each window. In the upper panels the solid black line is the test statistic, and the dotted red line is the 5% critical value. In the lower panels p-values are plotted with 0.05 denoted.
Figure 6: Max-Correlation Test in Rolling Window ($L_n = 8$)
Figure 6: Max-Correlation Test in Rolling Window ($L_n = 17$)
Figure 6: Max-Correlation Test in Rolling Window ($L_n = 21$)
Figure 6: Max-Correlation Test in Rolling Window ($L_n = 43$)

- Bands / Shanghai
- Bands / Nikkei
- Bands / FTSE
- Bands / SP500
- P-Values / Shanghai
- P-Values / Nikkei
- P-Values / FTSE
- P-Values / SP500
Andrews and Ploberger’s (1996) sup-LM test. Lag length is $L_n = \max\{5, \lceil \delta \times n/\ln(n) \rceil \}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$ and window size $n = 240$ trading days, hence $L_n = 5, 8, 17, 21, 43$. We also cover $L_n = n - 1 = 239$. We use the dependent wild bootstrap with 5,000 replications for each window. Block size is $b_n = \lceil c \sqrt{n} \rceil$, where $c \sim U(0.5, 1.5)$ is drawn independently across bootstrap samples and windows. Each point on the horizontal axis represents the initial date of each window. In the upper panels the solid black line is the test statistic, and the dotted red line is the 5% critical value. In the lower panels p-values are plotted with 0.05 denoted.
Figure 7: Andrews-Ploberger Sup-LM Test in Rolling Window ($L_n = 8$)
Figure 7: Andrews-Ploberger Sup-LM Test in Rolling Window ($L_n = 17$)
Figure 7: Andrews-Ploberger Sup-LM Test in Rolling Window ($L_n = 21$)
Figure 7: Andrews-Ploberger Sup-LM Test in Rolling Window ($\mathcal{L}_n = 43$)
Figure 7: Andrews-Ploberger Sup-LM Test in Rolling Window ($\mathcal{L}_m = 239$)
Shao’s (2011) Cramér-von Mises test. Lag length is \( \mathcal{L}_n = n - 1 = 239 \). We use the dependent wild bootstrap with 5,000 replications for each window. Block size is \( b_n = \lceil c \sqrt{n} \rceil \), where \( c \sim U(0.5, 1.5) \) is drawn independently across bootstrap samples and windows. Each point on the horizontal axis represents the initial date of each window. In the upper panels the solid black line is the test statistic, and the dotted red line is the 5% critical value. In the lower panels p-values are plotted with 0.05 denoted.